This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2014.2316151, IEEE Transactions on Affective Computing

Vision and Attention Theory Based Sampling for Continuous Facial Emotion Recognition

Albert C. Cruz, *Student Member, IEEE,* Bir Bhanu, *Fellow, IEEE,* and Ninad S. Thakoor, *Member, IEEE*

Abstract—Affective computing—the emergent field in which computers detect emotions and project appropriate expressions of their own—has reached a bottleneck where algorithms are not able to infer a person's emotions from natural and spontaneous facial expressions captured in video. While the field of emotion recognition has seen many advances in the past decade, a facial emotion recognition approach has not yet been revealed which performs well in unconstrained settings. In this paper, we propose a principled method which addresses the temporal dynamics of facial emotions and expressions in video with a sampling approach inspired from human perceptual psychology. We test the efficacy of the method on the Audio/Visual Emotion Challenge 2011 and 2012, Cohn-Kanade and the MMI Facial Expression Database. The method shows an average improvement of 9.8% over the baseline for weighted accuracy on the Audio/Visual Emotion Challenge 2011 video-based frame-level subchallenge testing set.

Index Terms—Facial expressions, Audio/Visual Emotion Challenge, Sampling and Interpolation

1 Introduction

FACIAL emotion recognition has applications in human-computer interaction, medical, advertising, and action recognition for computer games.

An emergent application of *Affective Computing* incorporates facial emotion and expression recognition. An embodied agent senses a person's emotion and projects an appropriate expression in response [1]. This facilitates non-verbal communication between a person and a computer, thus, improving feedback between them. However, state-of-the-art algorithms do not generalize to unconstrained data, presenting a challenge to this field.

Current methods perform well on datasets acquired in controlled situations, e.g. the Japanese Female Facial Expression database [2], Cohn-Kanade (CK) [3], the MMI Facial Expression Database (MMI-DB) [4], and the Facial Emotion Recognition and Analysis (FERA) challenge dataset [5]. However, the Audio/Visual Emotion Challenge (AVEC) datasets [6], [7] present difficult challenges. With previous datasets, each dataset was small enough to be loaded into memory at once, even for cases of high feature dimensionality. Previous approaches could reduce the number of frames to be processed by taking advantage of apexes of emotions, such as in CK. The most intense and discriminative frames corresponding to the apexes were labeled so a method could choose to retain them only.

The AVEC datasets explore the problems of a continuous emotion dataset, where it is computationally undesirable to select all the frames for processing. There are approximately one and a half million frames of video. The expressions in the dataset are subtle, spontaneous, and difficult to detect. The people in the videos are expressing emotions in a natural setting. The videos are not segmented. The apex labels are not given and it may be difficult to detect them automatically. In this paper, we propose a principled method for downsampling the frames for facial emotion and expression recognition. The method is inspired by the behavior of the human visual system. It can take advantage apexes if they are provided, but they are not required.

The rest of the paper is organized as follows: Section 2 discusses related work, motivations and contributions. Section 3 details the proposed downsampling method, and the full emotion recognition pipeline. Section 4 provides dataset information, parameters and results on AVEC 2011, AVEC 2012, CK and MMI-DB. Section 5 presents the conclusions of the paper.

2 Motivation, Related Work and Contributions

The motivation for sampling and reducing memory cost in large datasets is given in Section 2.1. A survey of related work, entries to the AVEC datasets, and other downsampling methods is given in Section 2.2. The contributions of this paper are given in Section 2.3.

- *A. C. Cruz, B. Bhanu and N. S. Thakoor are with the Center for Research in Intelligent Systems, University of California, Riverside, Winston Chung Hall 216, Riverside, CA, 92521-0425, USA.*
  E-mail: {acruz,bhanu,ninadst}@ee.ucr.edu.
  Acknowledgements to be added later.

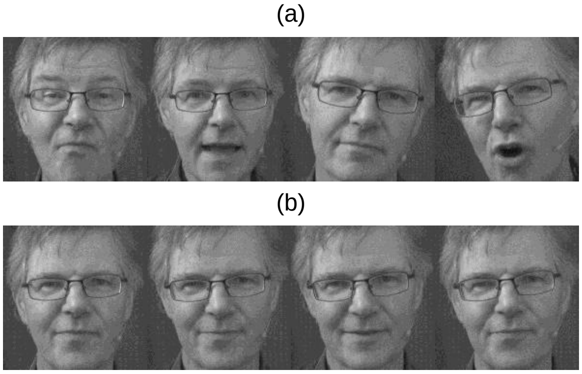


Fig. 1: Two different segments of AVEC [6] development video 14. (a) Many frames are required to describe the person's pose change and facial expressions. (b) The person is less expressive, and the segment needs few frames to be described.

2.1 Motivation

In the AVEC datasets, videos are captured at a high frame rate and over a long period of time. This makes it difficult to train a model for classification using all the frames in the dataset. An easy solution is to temporally downsample the video at a uniform, low frame rate. Unfortunately, this procedure results in a loss of precision, as it does not have the ability to precisely detect when the emotion changes. A dynamic sampling rate is desired that assigns a lower frame rate to parts of the video where the person is idle, and a higher frame rate to parts of the video where the person is animated. For example, in Figure 1, there are two different segments of the same video which merit different sampling rates. In Figure 1(a), the person is changing his pose, opening his mouth, furrowing his brow, using his cheek muscles, and raising his eyebrows. Many frames are needed to describe this segment. In Figure 1(b), the person holds his expression, so this segment would need only a few frames to be described. Therefore, we propose a method that applies a *dynamic sampling rate* which would allocate less frames for data analysis when the individual is idle, and more when the individual is active. The large volume of data poses the following problems to a downsampling procedure:

(1) With the AVEC datasets, processing each frame would be too costly. The downsampling should occur as early as possible in the video processing pipeline. Though related work [8], [9] propose dynamic downsampling, these methods prune samples *late* in the recognition pipeline, in classification.

(2) Use of the apex label is popular in facial expression and emotion recognition, and results show that features from the apex region improve classification rates [10]–[12]. However, the apexes must be *manually* labeled by an expert. If an algorithm is used to detect the apexes, the labeling can have errors. Situations may arise in the AVEC datasets where expressions are so subtle that extracting apex information is a difficult task for both humans and computers. There is a need for annotation free facial emotion and expression recognition. Our method does not require apex labels.

2.2 Related Work

In the baseline visual system for FERA [5] and the AVEC datasets [6], [7], face region-of-interest (ROI) is extracted which is then aligned by eye corner points. Subsequently, Local Binary Patterns (LBP) [26] are extracted as histogram-based features, and the emotions are classified with a support vector machine (SVM). In [24], the top approach for discrete emotions on the FERA dataset, Yang and Bhanu introduced a novel registration procedure called avatar image registration. It was found that a better registration method greatly improved performance. In [23], Valstar et al. tracked 20 fiducial facial points and classified them using a probabilistic actively learned SVM.

AVEC 2011 challenge [6]: In [20], Ramirez et al. quantified eye gaze, smile and head tilt with a commercial software (Omron OKAO Vision and Fraunhofer Sophisticated High-speed Object Recognition Engine) and used a Latent-Dynamic Conditional Random Field (LDCRF) [27] classifier. In [13], Glodek et al. modelled their system after the human perception's capability to separate form and motion. Gabor filters captured spatial information, and correlation features captured temporal information. The features were fed into multiple stages of filtering and non-linear pooling to further simulate human perception. In [8], Dahmane and Meunier proposed an approach for representation of the response to a bank of Gabor energy filters with histograms. A SVM with a radial basis function was used as a classifier.

AVEC 2012 challenge [7]: In [18], Nicolle et al. used 3-D model fitting, and global and local patch-based appearance features. These features were extended temporally with log-magnitude Fourier spectrum. A correlation based feature selector was proposed and a Nadaraya-Watson estimator was used as a classifier. During ground-truth labelling, the expert watches the video, and then notes changes in the label. There is a time delay between the actions in the video, and when the expert notes the change. Their method accounted for this delay. In [22], Soladi et al. employed two active appearance models, one to quantify head pose, and one to quantify smile. A *Mamdani* type fuzzy inference system was used. The features included who the person was speaking with, duration of sentences, and how well engaged the person was in the conversation with the embodied agent. In [16], Maaten used the baseline features, the derivative of features, and L_2-regularized linear least-squares regression. In [19], Ozkan et al. proposed a concatenated hidden Markov model (co-HMM). The label intensity values were discretized into bins. A HMM was trained to detect a specific bin, e.g., if there were ten quantization levels,

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2014.2316151, IEEE Transactions on Affective Computing

TABLE 1: Review of Related Work. AAM: active appearance model. AIR: avatar image registration. CRF: conditional random field. HMM: hidden Markov model. LBP: local binary patterns. LLS: linear least squares. LPQ: local phase quantization. MHI: motion history images. SVM: support vector machine.

Approach	Downsampling	Registration	Features	Classifier	Dataset
AVEC 2012 Baseline [7]	Fusion of 50 frames	Eye-point	LBP	SVM	AVEC 2012 [7]
Dahmane and Meunier [8]	Change granularity if label changes	Eye-point[1]	Histograms of Gabor	SVM	AVEC 2011 [6]
Glodek et al. [13]	Random	Eye-point[1]	Gabor, temporal correlation	SVM, HMM	AVEC 2011 [6]
Jiang et al. [14]	Random, bootstrapping, heuristic	Eye-point	LPQ from Three Orthogonal Planes	SVM, modelling of temporal phases	FERA [5], MMI-DB [4], SAL, UNBC-McMaster pain
Koelstra et al. [15]	X	Affine	MHI, orientation histograms	Gentleboost, HMM	MMI-DB [4], CK [3]
Maaten [16]	X	Eye-point[1]	LBP	LLS	AVEC 2012 [7]
Meng and Bianchi-Berthouze [17]	X	Eye-point[1]	LBP	Multi-HMM	AVEC 2011 [6]
Nicolle et al. [18]	X	Point distribution model	Eigenappearance, log-magnitude Fourier spectra	Nadaraya-Watson	AVEC 2012 [7]
Ozkan et al. [19]	X	Commercial	Commercial, frame number	Level quantization, co-HMM	AVEC 2012 [7]
Ramirez et al. [20]	X	Commercial	Commercial	Latent-dynamic CRF	AVEC 2011 [6]
Savran et. al. [21]	Select outlier frames based on standard deviation	Eye-point[1]	Local appearance statistics	Bayesian filtering fusion	AVEC 2012 [6]
Soladi et al. [22]	X	AAM	Statistics of head pose	Fuzzy inference system	AVEC 2012 [6]
Valstar et al. [23]	X	Particle filtering with factorized likelihoods	Fiducial facial points	Probabilistic active learning SVM	MMI-DB [4], CK [3]
Wu et al. [10]	X	None stated	Spatiotemporal Gabor	Bootstrapping, SVM	CK [3]
Yang and Bhanu [24]	X	AIR	LBP and LPQ	SVM	FERA [5]
Zhu et al. [9]	Bootstrapping to select frames based on apexes	AAM, eye-point	Tracker points, SIFT	AdaBoost	RU-FACS [25]
Proposed Method	Dynamic sampling based on changes in visual information with or without apex	AIR	LBP	SVM	AVEC 2011/2012 [6], [7], MMI-DB [4], CK [3]

then there would be ten classifiers each detecting if that specific level was present. A final HMM fused these outputs at the decision level. In the video-based approach in [21], Savran et al. extended local appearance features to the temporal domain by taking the mean and standard deviation in sliding temporal windows. AdaBoost was used a feature selector, and ϵ-support vector regression (SVR) was used to regress the labels.

Sampling methods: Some approaches have attempted to address the sampling issue. In [13], Glodek randomly sampled the video frames. In [8], a downsampling method was proposed that changed granularity of sampling based on whether or not a change was detected in the predicted label. A limitation of this system is that it assumes that the system can correctly predict the label. In [9], Zhu et al. reduced the number of frames in the dataset with a bootstrapping procedure. This method requires the apexes to be labeled. We propose a method that does not require peak frame labeling. In [21], Savran et al. downsampled the training data to frames that had an emotion label intensity greater than $\pm\sigma$ from the mean emotion intensity. No framework for downsampling test data was provided. In [14], Jiang et al. proposed a texture descriptor that extended Local Phase Quantization (LPQ) features to the temporal domain. It was called Local Phase Quantization from Three Orthogonal Planes. The paper also investigated three downsampling methods: randomly selecting frames, bootstrapping, and a heuristic approach that found two subsets of the data to describe static appearance descriptors and dynamic appearance descriptors. It was found that the heuristic method was the best performer. All of these methods have focused on training data selection, and no method was given to downsample the testing data. A summary of related work is given in Table 1. As compared to the previous related work, the contributions of this paper are given below.

2.3 Contributions

We propose emulating the behavior of the human visual system to address the challenges in the AVEC datasets. The focus of work in this paper is video-based temporal sampling. The contributions of this paper are: (1) We exploit vision and attention theory [28], [29] from perceptual psychology to determine an appropriate sampling rate. We assign a dynamic, temporal granularity that is inversely proportional to how frequent the visual information on a person's face is changing. The method improves average correlation with the ground-truth for all affect dimensions on the AVEC 2012 frame-level subchallenge testing set

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2014.2316151, IEEE Transactions on Affective Computing

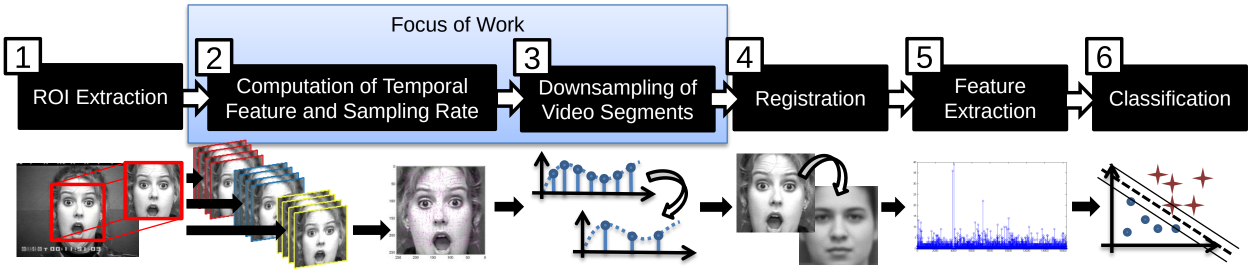


Fig. 2: System overview. (1) Extraction of ROI. (2) Partitioning of video into smaller segments, formation of temporal feature that quantifies motion, and computation of the dominant frequency of the temporal feature. (3) Downsampling of the video segment. (4) Registration of frames. (5) Appearance feature extraction. (6) Classification/regression.

over the baseline approach by a factor of 2.7. (2) We provide a framework for the method to integrate information from apex labels, if they are provided. The method improves average F_1 measure across 14 different classes by 7.6 over [24]. (3) We provide a framework for using match-score fusion temporally. The method improves average weighted accuracy on all classes on the AVEC 2011 frame-level subchallenge development set over the use of uniform sampling of 1 frame per segment and no fusion by 5.4%.

3 Technical Approach

When viewing a natural scene, the human visual system exhibits a saccade-fixation-saccade pattern [30]. *Fixations* are moments of attention, where visual information is being processed. *Saccades* are rapid movements of eyes, where information is not being processed. First the eyes saccade, then fixate, and this procedure is repeated. The latency between two saccades decreases with the increasing frequency of temporal changes of visual information in the scene. We propose a method that emulates this process for emotion and expression recognition. Human perception of faces is different than recognition of scenes or other objects. However, the focus of work is the concept of *attention*, the length of focus on a scene, not recognition. The temporal frequency of visual information in the scene affects the amount of attention given to a part of the scene. Our algorithm is inspired by this physical process and emulates attention by downsampling a video.

The overview of this work is shown in Figure 2: (1) face ROI is detected with Viola-Jones [31]. (2) The video is partitioned into segments. Within each segment, the visual information is quantified with temporal features. We apply a discrete Fourier transform to the temporal feature to find the *dominant frequency*, the frequency of the temporal feature with the most energy. (3) The video is downsampled at the dominant frequency. (4) The selected frames after the downsampling are aligned with avatar image registration [24]. (5) Appearance features are generated in local regions. (6) Initial *a posteriori* probabilities of emotion labels in each frame in the video segment are generated from SVM [32]. The results are temporally fused at the match-score level [33] to generate the final predicted labels. Section 3.1 discusses downsampling for continuous videos, Section 3.2 discusses downsampling when apex labels are given. The full emotion recognition pipeline is described in Section 3.3.

3.1 Downsampling Continuous Video

Downsampling of a continuous video without time annotations for apexes is done as data comes in. The videos are segmented into uniformly sized smaller segments. Each segment is downsampled *dynamically*, and each segment has its own appropriate downsampling factor. Conventionally, each segment would be processed with a *uniform* downsampling factor. Psuedocode for the downsampling method is given in Algorithm 1.

3.1.1 Time partitioning procedure

The video I is segmented into equally sized non-overlapping segments of N frames. The segment of video I_Φ contains the frames at indices Φ where $\Phi = \{m_0, m_0+1, ..., m_0+N-1\}$. The downsampled video segment I_{Φ^*} contains the frames at indices Φ^*, where Φ^* is a subsequence of Φ. Initially, the system delays for N frames, and processes a video segment of N frames at a time. We start with $m_0 = 0$, so the first N frames form one segment. Then $m_0 = N$, so the frames from N to $2N-1$ form another segment and so on, until the end of the video. If there is a remainder, it forms its own segment. We chose parameter N such that the duration of each segment is 1 s because 1 Hz is the maximum bound of the HVS according to vision and attention theory [30].

3.1.2 Computing the temporal feature

I_{Φ^*} is created by resampling I_Φ at a lower frequency. The first step is to quantify facial expressions into a signal that varies with time. The signal's frequency must respond to changes of facial expression. Because the frame rate is high, and the ROI is a frontal face, optical flow can be exploited to quantify the facial expressions [34]. ΔI_n is optical flow between the

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2014.2316151, IEEE Transactions on Affective Computing

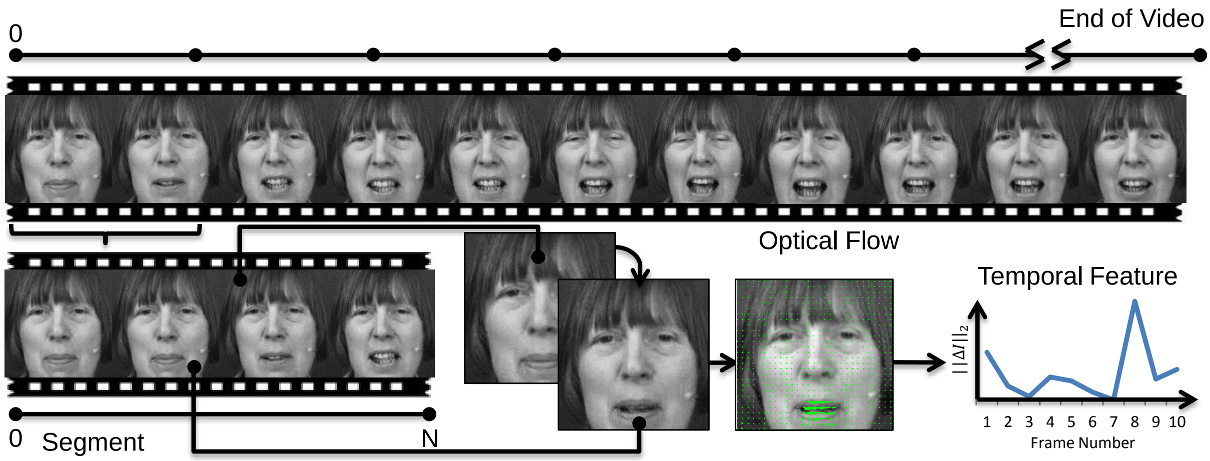


Fig. 3: Overview of how the temporal feature is computed. The video is segmented into non-overlapping segments of length N. Optical flow is computed using a pair of adjacent frames. The result of the optical flow forms the temporal feature.

frames I_n and I_{n-1}. It outputs a motion vector. The magnitude is summed for all pixels in an image to form a 1-D signal:

$$f\left(n\right) = \sum_{\mathbf{x}} \|\Delta I_n\left(\mathbf{x}\right)\|_2 \tag{1}$$

where $f\left(n\right)$ is the temporal feature for a single frame, $\mathbf{x}$ is a pixel, and $\|.\|_2$ is the magnitude. For the entire segment I_Φ, the temporal feature $\mathbf{f}_\Phi$ is indicated by: $\mathbf{f}_\Phi \equiv \left[f\left(m_0\right), f\left(m_0+1\right), ..., f\left(m_0+N-1\right)\right]$. Figure 3 shows how the video is segmented, how the optical flow is computed, and how the temporal feature is generated. As registration is costly, to reduce the number of frames to be registered, we compute the optical flow before registration. We do not use optical flow as a feature for classification, or for alignment.

3.1.3 Downsampling the video segment

To compute the dominant frequency, first, the DC-offset is removed:

$$\tilde{\mathbf{f}}_\Phi = \mathbf{f}_\Phi - \mathrm{E}\left(\mathbf{f}_\Phi\right) \tag{2}$$

where $\mathrm{E}\left(.\right)$ is the mean. It is important to remove the DC-offset for two reasons: (1) it normalizes the temporal feature and (2) for real data, the $\mathbf{F}_\Phi\left(0\right)$–corresponding to the coefficient at 0 Hz, the DC-offset–will be greater than other values of $\mathbf{F}_\Phi$, causing it to be selected as the dominant frequency. $\mathbf{F}_\Phi$ is the discrete Fourier transform of $\tilde{\mathbf{f}}_\Phi$: $\mathbf{F}_\Phi = \mathrm{DFT}\left(\tilde{\mathbf{f}}_\Phi\right)$, where $\mathrm{DFT}\left(.\right)$ is the discrete Fourier transform, and k is the frequency index. The frequency index corresponding to the frequency with the most energy β is computed as follows:

$$\beta = \mathrm{argmax}_k \|\mathbf{F}_\Phi(k)\| \tag{3}$$

where $\|\mathbf{F}_\Phi(k)\|$ is the magnitude of $\mathbf{F}_\Phi(k)$. Note that the frequency in Equation (3) is not the Nyquist rate. The Nyquist rate applies to sampling a continuous signal in order to accurately reconstruct that signal.

Algorithm 1 Computing the sampling rate for single segment/single apex

Input: I_Φ, the video segment. n_0, midpoint-apex time point (if given). N, number of frames in Φ.
Output: I_{Φ^*}, downsampled video segment.

```
procedure DOWNSAMPLESEGMENT(I_Φ)
    for all frames n ∈ Φ do
        ΔI_n ← optical flow from n − 1 to n
        f(n) = Σ_x ||ΔI_n(x)||_2
    end for
    f_Φ ← vector corresponding to all features f
    f̃_Φ ← f_Φ − mean of f_Φ
    F_Φ ← Discrete Fourier transform of f̃_Φ
    β ← argmax_k ||F_Φ(k)||
    if n_0 is given then
        Φ*_Apex ← range n_0 − β/2 < n ≤ n_0 + β/2
        Φ* ← Φ*_Apex
    else
        M ← N/β                              ▷ (Downsampling factor)
        Φ* ← Φ ↓ M                           ▷ (Every M-th frame)
    end if
    return I_Φ*
end procedure
```

In this paper we are downsampling a discrete signal by removing samples in the signal which have not changed much. For this reason, we sample at the dominant frequency itself.

The downsampling factor M is given by: (maximum frequency/dominant frequency). The frequency index β can be converted to the dominant frequency as: $2\pi\beta/N$. The maximum frequency index N corresponds to frequency 2π. It follows that: $M = N/\beta$. Let $\Phi^* = \Phi \downarrow M$. That is, Φ^* is every M-th frame of Φ. When the temporal feature has a high frequency, $\beta \rightarrow N$, the downsampling factor is near 1, and all of the frames are preserved. When the temporal feature has a low frequency, the downsampling factor increases, and most of the frames are removed.

3.2 Downsampling with Apex Labels

When apex label information is given, instead of segmenting the video evenly, the system segments the

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2014.2316151, IEEE Transactions on Affective Computing

Fig. 4: Comparison of sampling at even intervals versus sampling at the apex. A video is given, and its expression intensity is given. Sampling at even intervals retains frames that are further away from the apex. They are weakly expressed, and they are not a good representation of the emotion being expressed. Sampling at the apex retains the frames where the emotion is most strongly expressed.

video into durations centered at each apex. Instead of downsampling the segment evenly, the dominant frequency effects the duration of the segment. If the dominant frequency is high, then the method will select many frames at the apex; if low, only the frames nearest to the apex are selected. The human visual system has dynamic attention based on the changes of visual information. We realize attention as the number of selected frames. If there is not much change in the visual information, there is less attention given, and fewer frames are selected.

3.2.1 Time partitioning procedure

If apexes are provided, the video is partitioned into uniform segments of N frames, centered at the mid-point of the apex frames. There is a segment for each apex, and each segment is centered at that apex. Frames that are not near an apex will be removed. Let n_0 be the location of an apex. It now follows that:

$$\Phi_{\mathtt{Apex}} = \{n : n_0 - N/2 < n \leq n_0 + N/2\} \qquad (4)$$

Ordinarily we downsample the segment evenly. However, when apex labels are given we reformulate the downsampling method to take advantage of these labels. At the apex, the expressions are strong and the emotion is more easily detected. For this reason, the frames in the duration centered at the apex should be retained, rather than downsampling uniformly, which may retain frames further away from the apex where emotions are more difficult to detect. An example comparing sampling at a uniform rate versus sampling at the apex is given in Figure 4. There is no change in the way β is computed.

3.2.2 Downsampling the video segment

In this formulation, $\Phi^*_{\mathtt{Apex}}$ varies in duration according to β, and is defined as follows:

$$\Phi^*_{\mathtt{Apex}} = \{n : n_0 - \beta/2 < n \leq n_0 + \beta/2\} \qquad (5)$$

If apex labels are given, $\Phi^*_{\mathtt{Apex}}$ is taken to be Φ^*. When the temporal feature has a high frequency, N frames are preserved and I_{Φ^*} is equivalent to $I_{\Phi_{\mathtt{Apex}}}$. When the feature has a low frequency, the number of frames approaches 1, and most of the frames are removed.

3.3 Emotion Recognition System Pipeline

3.3.1 Face ROI extraction, registration and features

Faces are detected with a boosted cascade of Haar-like features [31]. If a face is not detected in the frame, we assign the expected label to that frame. For classification, we assign the class label that has the highest percentage of class occurrence. For regression, we assign the average value of the emotion intensity from the training data. A better method for assigning the label in this situation would be a first-order Markov assumption, but this is not the focus of work (see [35]). If ROI is detected, faces are registered with avatar image registration. The reader is referred to [24] for a more in depth explanation. We use Local Binary Patterns (LBP) because they are the most popular features in the field for representing a face. The reader is referred to [36], [37] for an in depth explanation. The features are computed for each frame in I_{Φ^*}.

3.3.2 Fusion

A method is needed to temporally fuse and smooth the estimated emotions. For each segment I_{Φ^*}, we propose fusing the *a posteriori* probabilities for each frame computed by the classifier. *A posteriori* probabilities are obtained with SVM [32]. The *a posteriori* probabilities are fused with combination-based match-score fusion [33], in which the scores, or *a posteriori* probabilities, from different matchers are weighted and combined to obtain a final, single score as the *a posteriori* probability. Let $\mathbf{y}_j$ be the feature vector of LBP features of frame j in I_{Φ^*}. c_i is the class label from one of the classes: $c_1, ..., c_{n_c}$. The estimated label for all the frames in I_Φ is $\tilde{c}$. Note that this assigns labels to all frames Φ, including those that were not selected for processing. Temporal smoothing is introduced by assigning all the frames in I_{Φ^*} the same label. $p\left(c_i|\mathbf{y}_j\right)$ is the *a posteriori* probability of a class c_i. The first step

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2014.2316151, IEEE Transactions on Affective Computing

TABLE 2: Percentage of positively expressed affective dimension for the AVEC 2011 video sub-challenge.

Sets	Arousal	Expectancy	Power	Valence
Training	47	46	51	55
Develop	56	40	59	64

TABLE 3: Percentage of positively expressed AU for CK.

AU1	AU2	AU4	AU5	AU6	AU7	AU9
29.2	19.6	31.7	16.0	22.7	22.1	10.2

AU10	AU12	AU15	AU20	AU24	AU25	AU27
2.5	23.1	15.1	14.1	8.6	60.1	15.5

of fusion is estimation of $p\left(c_i|\mathbf{y}_j\right)$ for each frame in I_{Φ^*} with the method in [38].

The second step aggregates the *a posteriori* probabilities from the selected frames into a single score. The classification rule for match-score fusion is:

$$\tilde{c} = \text{argmax}_{c_i} h\left(c_i, \Phi^*, \mathbf{y}_1, ..., \mathbf{y}_{n_f}\right) \tag{6}$$

where $h(.)$ is the rule for aggregation, and n_f is the number of frames in Φ^*. The *Sum rule* is as follows:

$$h_{\texttt{Sum}}\left(c_i, \Phi^*, \mathbf{y}_1, ..., \mathbf{y}_{n_f}\right) = \frac{1}{n_f}\sum_{j\in\Phi^*} p\left(c_i|\mathbf{y}_j\right) \tag{7}$$

The *Product rule* is as follows:

$$h_{\texttt{Product}}\left(c_i, \Phi^*, \mathbf{y}_1, ..., \mathbf{y}_{n_f}\right) = \prod_{j\in\Phi^*} p\left(c_i|\mathbf{y}_j\right) \tag{8}$$

The *Min and Max rules* are as follows:

$$h_{\texttt{Min}}\left(c_i, \Phi^*, \mathbf{y}_1, ..., \mathbf{y}_{n_f}\right) = \min_{j\in\Phi^*} p\left(c_i|\mathbf{y}_j\right) \tag{9}$$

$$h_{\texttt{Max}}\left(c_i, \Phi^*, \mathbf{y}_1, ..., \mathbf{y}_{n_f}\right) = \max_{j\in\Phi^*} p\left(c_i|\mathbf{y}_j\right) \tag{10}$$

The *Mode rule* $h_{\texttt{Mode}}$, differs from the above rules by assigning the most common label to each frame in the segment.

The approach can be applied to regression by taking the result of the aggregation rule to be the final decision value. This replaces Equation 6, where a second classifier is applied:

$$\tilde{c}_{\texttt{Regression}} = h\left(c_i, \Phi^*, \mathbf{y}_1, ..., \mathbf{y}_{n_f}\right) \tag{11}$$

Note that, for regression, we do not estimate the *a posteriori* probability. $p(.)$ in the above equations is replaced with the decision values from SVR [32].

4 EXPERIMENTS

4.1 Datasets

AVEC 2011 [6] and 2012 [7] are grand challenge datasets. In this paper, they are used to compare the proposed method to other state-of-the-art methods. It is a non-trivial, unconstrained dataset: (1) the frame rate is too high to load all frames into memory. For example, if AVEC 2012 has 1351129 frames, if LBP features and baseline audio features [6] are used which have 7841 dimensions, and if double floating points are used for each feature, it would require 8.48 GB to load all frames into memory. This exceeds the memory of most computers (88.9% of computers have up to only 8 GB of computer memory according to a recent hardware survey [39]). (2) The subjects are free to change pose, and use hand gestures, and (3) the videos are not acted. The videos are not pre-cut, and a person can express multiple emotions per video. In the AVEC datasets, a person is presented with the Sensitive Artificial Listener [40] who engages the person in conversation, and causes emotionally colored conversations by being biased to express a particular emotion, such as belligerence or sadness. Emotions expressed in this scenario are natural, continuous, and spontaneous. An example is available online [41]. In this example, a person is interacting with a specific character named Spike. Spike is confrontational, and aggravates the person during conversation. Note that the person is smiling, but not from being pleased. The smile is caused by the person being polite and exercising restraint in response to hostility. A separate classifier is used for each affect dimension (see Section 4.2).

TABLE 4: Percentage of classes for MMI-DB emotions.

Anger	Disgust	Fear	Happy	Sad	Surprise
21.1	13.9	13.0	19.7	14.4	17.9

The AVEC datasets are divided into three partitions: (a) 31 interviews of 8 different individuals form the *training* set. It is used as samples for a training model. (b) 32 interviews of 8 individuals, who are different from the training set form the *development* set. It is used as the testing fold in the training phase, and (c) 32 (AVEC 2012) or 11 (AVEC 2011) interviews of new individuals who are not in the development or training set form the *testing* set. The testing set is the official validation fold with which algorithms are compared to each other. The average length of all the videos in AVEC 2011 is $14.6\times10^3 \pm 5.20\times10^3$ frames. All results are given in terms of the frame level subchallenge. The percentage of positively expressed affective dimension for the training and development datasets for AVEC 2011 dataset are given in Table 2. The percentages for the testing set are not available because the labels are withheld by the challenge organizers.

The second dataset used is CK [3]. We use this database to test the quality of results of the proposed sampling method, when apex labels are provided. The length of segments range from 3 frames to over 100. The percent of positively expressed AU are given in Table 3. We follow the testing methodology in Koelstra et al. [15]. An AU is selected if it has more than 10 positive examples. We detect the following actions units (AU): $\{1, 2, 4, 5, 6, 7, 9, 10, 12, 15, 20, 24, 25, 27\}$. The reader is referred to Lucey et al. [3] for a more detailed expla-

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2014.2316151, IEEE Transactions on Affective Computing

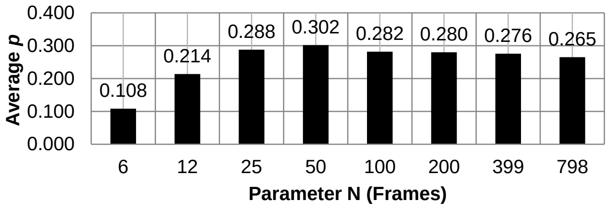


Fig. 5: Average correlation of all affect dimensions on development set, AVEC 2012 frame-level subchallenge for varying values of N.

nation of the data. We use leave-one-person-out cross-validation. A binary classifier is used for each AU.

MMI-DB [4] is frontal face video data similar to CK. For most videos, the emotion peaks near the middle of the video. The percentage of class for each emotion is given in Table 4. We use leave-one-person-out cross validation. We use all sessions that have emotion labels, and we consider the classes with at least 10 positive examples. We use only frontal faces. A multi-class classifier is used.

4.2 Expression and Emotion Labels

We use three labeling systems: action units [42], emotions based on the Ekman big six [42] and the Fontaine emotional model [43]. Expressions and emotions are not the same. Expressions are facial muscle movements. Ekman and Friesen [42] defined the minimal set of facial muscle movements, or action units (AUs), that are used in expressions. This is the Facial Action Coding System. Emotion differ from expressions in that they are the underlying mental states that may illicit expressions. A common system for discrete emotional states is the Ekman big six: happiness, sadness, fear, surprise, anger and disgust.

A different system for emotion labels is the Fontaine emotional model [43] with four affect dimensions: *valence*, *arousal*, *power* and *expectancy*. An emotion occupies a point in this four-dimensional Euclidean space. Valence, also known as evaluation-pleasantness, describes positivity or negativity of the person's feelings or feelings of situation, e.g., happiness versus sadness. Arousal, also known as activation-arousal, describes a person's interest in the situation, e.g., eagerness versus anxiety. Power, also known as potency-control, describes a person's feeling of control or weakness within the situation, e.g., power versus submission. Expectancy, also known as unpredictability, describes the person's certainty of the situation, e.g., familiarity versus apprehension. For a more detailed explanation, the reader is referred to [43]. With this system, multiple emotions can be expressed at the same time. An Ekman big six emotion [44] occupies a point in each of these four dimensions.

An expression or emotion also has intensity. It can be continuous, where the label has a numerical value representing its intensity, such as in AVEC 2012 [7]. The intensity can also be discrete, where the numerical values have been categorized into bins. In CK [3], an AU is either expressed (positive) or not expressed (negative). In AVEC 2011, the intensity was quantized into values higher than the average value (positive), or lower than the average value (negative). We use discrete action units for CK, discrete big six-based emotions for MMI-DB, discrete Fontaine for AVEC 2011 and continuous Fontaine for AVEC 2012.

Another system for level quantization has four states: neutral, onset, apex and offset [15]. These states indicate the intensity of an emotion, e.g., an expression is neutral when it has no expression, and an expression is at its apex when it has its greatest intensity. These four states form a state space. A person's expression will transition between these states, e.g. over time it will go from neutral to onset to apex.

4.3 Performance Metrics

The AVEC datasets have two scoring systems. In AVEC 2011 [6] the metrics are weighted accuracy (WA) and unweighted accuracy (UA). Weighted accuracy is the classification rate, and is also known as percent correct, calculated as follows:

$$\mathrm{WA} = \frac{1}{n_c}\sum_{i=1}^{n_c} p\left(c_i\right)\frac{\mathrm{tp}^i}{\mathrm{tp}^i + \mathrm{fp}^i} \tag{12}$$

where tp_i is the number of true positives of class i, fp_i is the number of false positives of class i, and n_c is the number of classes and $p\left(c_i\right)$ is the percentage of class. Unweighted accuracy is defined as:

$$\mathrm{UA} = \frac{1}{n_c}\sum_{i=1}^{n_c}\frac{\mathrm{tp}_i}{\mathrm{tp}_i + \mathrm{fp}_i} \tag{13}$$

This metric is used because some classes in the data have disproportionate percentage. For example, positive valence has a percentage of class higher than 60% in the training fold. The results for AVEC 2012 are given in terms of the Pearson product-moment correlation coefficient with the ground-truth labels. It is computed as:

$$\rho = \frac{\mathrm{E}\left(\left(\mathbf{c} - E\left(\mathbf{c}\right)\right)\left(\tilde{\mathbf{c}} - E\left(\tilde{\mathbf{c}}\right)\right)\right)}{\sigma_{\mathbf{c}}\sigma_{\tilde{\mathbf{c}}}} \tag{14}$$

where $E\left(.\right)$ is the mean, $\mathbf{c}$ are the ground-truth labels across all persons and videos concatenated into a single vector. $\tilde{\mathbf{c}}$ are the estimated labels across all persons and videos concatenated into a single vector; $\mu_{\mathbf{c}}$ and $\mu_{\tilde{\mathbf{c}}}$ are the mean of the ground-truth and predicted labels, respectively; and $\sigma_{\mathbf{c}}$ and $\sigma_{\tilde{\mathbf{c}}}$ are the standard deviation of the ground-truth and predicted labels, respectively. CK comparisons are quantified with the F_1 measure [15]:

$$F_1 = 2\left(\frac{\left(\mathrm{precision}\right)\left(\mathrm{recall}\right)}{\mathrm{precision} + \mathrm{recall}}\right) \tag{15}$$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2014.2316151, IEEE Transactions on Affective Computing

TABLE 5: Weighted Accuracy Results for Various Sampling Methods, Registration Methods and Fusion Methods for AVEC 2011 Development set. Sampling: sampling rate. Uniform: uniform number of frames. Reg: registration method. AIR: avatar image registration. RST: similarity transform. Rule: fusion rule. HMM: hidden Markov model. WA: weighted accuracy.

			WA Result				
Sampling	**Reg**	**Rule**	**Aro**	**Exp**	**Pow**	**Val**	**Avg**
Proposed	AIR	Sum	71.7	62.1	63.4	65.3	65.6
Proposed	AIR	Max	71.0	60.7	63.2	64.8	64.9
Proposed	RST	Min	70.1	61.0	62.1	65.0	64.5
Proposed	RST	Mode	71.0	61.9	61.8	62.6	64.3
Proposed	RST	Sum	70.7	60.2	63.0	63.0	64.2
Proposed	RST	Prod	69.6	61.9	61.2	62.8	63.9
Proposed	RST	Max	69.0	60.1	61.6	64.6	63.8
Proposed	AIR	HMM	68.5	62.0	59.8	64.9	63.8
Proposed	AIR	Prod	70.2	59.8	60.5	64.3	63.7
Proposed	AIR	Mode	71.6	59.5	60.9	62.6	63.6
Proposed	RST	No	69.0	59.6	62.1	63.6	63.6
Proposed	AIR	Min	70.1	59.2	60.8	62.6	63.2
Proposed	AIR	No	69.1	55.5	62.5	64.7	62.9
Uniform 3	AIR	Sum	69.3	57.7	61.0	63.7	62.9
Uniform 6	AIR	Sum	67.7	60.0	57.9	62.9	62.1
Uniform 9	AIR	Sum	67.6	57.2	60.2	61.4	61.6
Uniform 6	AIR	Mode	67.9	56.7	58.7	62.3	61.4
Uniform 3	AIR	Mode	65.9	61.6	59.0	58.5	61.2
Uniform 9	AIR	Mode	68.3	55.6	58.8	58.6	60.3
Uniform 1	AIR	No	65.0	56.3	57.0	62.4	60.2

It is the harmonic mean of precision and recall. It can be more meaningful in cases of disproportionate percentage of different classes.

4.4 Parameters

After ROI extraction, all face images are resized to 200×200 with bicubic interpolation. For avatar image registration, we train the avatar reference image from the development data subsampled at 12 fps. The parameters specific to avatar image registration are: $\alpha = 2$, $1/(\sigma)^2 = .005$, and the number of iterations is 3. All three of these parameters are empirically selected from the previous work [24]. The parameters specific to LBP [26] are: the number of local regions is 8, patterns are computed for 8 neighbors at a radius of 1, and there are 10×10 sub-regions on the entire face image. All classifiers are SVM [32]. The parameters specific to the SVM are: an RBF kernel is used, the cost $c = 1$, and $\gamma = 2^{-8}$. The feature vectors are normalized to $[-1, 1]$. For regression, an ϵ-SVR is used [32]. The parameters specific to the regressor are: $\epsilon = 0.1$.

N is the initial number of frames. There should be enough frames in Φ to describe the expression in progress. In the unconstrained case, an expression can be very quick. If that expression were a microexpression, it could be as fast as $1/25$th of a second, requiring 25 fps [45]. MMI-DB videos were captured at 24 fps, so we recommend that $N > 24$ for MMI-DB. We chose $N = 50$ frames. It is validated empirically. AVEC 2012 is used for selecting parameter N. A value is selected empirically by varying N in powers of 2 seconds: $\{2^{-3}, ..., 2^{8}\}$. The results are given in Figure 5. $N = 50$ gives the best performance. It decreases as N is reduced below 50 frames. For decreasing values of N, the upper bound of β decreases, and more frames will be forced to be selected. The worst performer is 6 frames per segment.

TABLE 6: Confusion Matrices for MMI-DB. An: anger. Di: disgust. Fe: fear. Ha: happiness. Sa: sadness. Su: surprise.

(a)

Yang and Bhanu [24]

	An	Di	Fe	Ha	Sa	Su
An	71.7	2.2	2.2	6.5	4.4	13.0
Di	12.9	48.4	16.1	6.5	0.0	16.1
Fe	27.6	0.0	58.6	3.5	0.0	10.3
Ha	9.5	0.0	4.8	76.2	0.0	9.5
Sa	25.0	0.0	6.3	6.3	59.4	3.1
Su	18.4	2.6	7.9	0.0	5.6	65.8

(b)

Uniform Sampling of 1 Frame

	An	Di	Fe	Ha	Sa	Su
An	76.4	4.7	6.8	2.2	2.2	8.7
Di	9.7	64.5	9.7	3.2	3.2	9.7
Fe	24.1	0.0	55.2	0.0	6.9	13.8
Ha	11.9	0.0	2.4	76.2	2.4	7.1
Sa	28.1	0.0	6.3	3.1	53.1	9.4
Su	21.1	7.9	5.3	0.0	0.0	65.8

(b)

Proposed with Frame Differencing as Temporal Feature

	An	Di	Fe	Ha	Sa	Su
An	78.3	6.5	0.0	4.4	4.4	6.5
Di	9.7	67.7	12.9	0.0	0.0	9.7
Fe	27.6	0.0	58.6	3.5	0.0	10.3
Ha	14.3	7.1	9.5	61.9	0.0	7.1
Sa	21.9	0.0	6.3	0.0	62.5	9.4
Su	15.8	2.6	2.6	0.0	2.6	76.3

(c)

Proposed with Dense-SIFT as Temporal Feature

	An	Di	Fe	Ha	Sa	Su
An	76.1	6.5	0.0	0.0	4.4	13.0
Di	9.7	58.1	16.1	3.2	0.0	12.9
Fe	17.2	0.0	69.0	3.5	0.0	10.3
Ha	14.3	4.8	2.4	69.1	0.0	9.5
Sa	21.9	3.1	0.0	3.1	59.4	12.5
Su	18.4	0.0	2.6	0.0	0.0	79.0

(d)

Proposed with Optical Flow as Temporal Feature

	An	Di	Fe	Ha	Sa	Su
An	73.9	4.4	4.4	0.0	8.7	8.7
Di	6.5	74.2	6.5	0.0	0.0	12.9
Fe	17.2	3.5	69.0	0.0	0.0	10.3
Ha	9.5	4.8	2.4	76.2	0.0	7.1
Sa	21.9	0.0	0.0	3.1	71.9	3.1
Su	21.1	2.6	5.3	2.6	2.6	65.8

TABLE 7: Weighted accuracy and unweighted accuracy on MMI-DB for varying temporal features. Prop.: Proposed. UA: unweighted accuracy. WA: weighted accuracy.

Method	**WA**	**UA**
Yang and Bhanu [24]	63.4	64.8
Uniform Sampling of 1 Frame	65.2	66.6
Prop. + Frame Differencing Temporal Feature	67.6	68.4
Prop. + Dense-SIFT Temporal Features	68.4	69.4
Prop. + Optical Flow Temporal Feature	71.8	72.0

4.5 Experimental Results

Training results that select the best performing combination of registration method and fusion rule are

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2014.2316151, IEEE Transactions on Affective Computing

TABLE 8: Comparison to Other Methods on AVEC 2011 Frame-level Subchallenge Testing Set. Bold indicates best performer, underline indicates second best.

	(a) Development Set									
	Arousal		Expectancy		Power		Valence		Average	
Method	WA	UA	WA	UA	WA	UA	WA	UA	WA	UA
Proposed Method	**71.7**	**67.8**	**62.1**	**59.8**	**63.4**	**61.8**	**65.3**	60.7	**65.6**	**62.6**
Glodek et al. [13]	58.2	53.5	53.6	53.2	53.7	53.8	53.2	49.8	54.7	52.6
Dahmane and Meunier [8]	54.9	55.0	51.8	51.2	53.2	52.8	56.6	55.5	46.6	53.6
Baseline [6]	60.2	57.9	58.3	56.7	56.0	52.8	63.6	**60.9**	59.5	57.1

	(b) Testing Set									
	Arousal		Expectancy		Power		Valence		Average	
Method	WA	UA	WA	UA	WA	UA	WA	UA	WA	UA
Proposed Method	56.5	56.9	**59.7**	**55.1**	**48.5**	**49.4**	**59.2**	**56.7**	**56.0**	**54.5**
Glodek et al. [13]	56.9	57.2	47.5	47.8	47.3	47.2	55.6	55.6	51.8	52.0
Dahmane and Meunier [8]	**63.4**	**63.7**	35.9	36.6	41.4	41.1	53.4	53.6	48.5	48.8
Baseline [6]	42.2	52.5	53.6	49.3	36.4	37.0	52.5	51.2	46.2	47.5

TABLE 9: Comparison to Other Methods on AVEC 2012 Video-based Frame-level Subchallenge Testing and Development Sets. Bold indicates best performer, underline indicates second best. Aro: arousal. Exp: expectancy. Pow: power. Val: valence. Avg: average of all.

Video-only Development Set					
Method	Aro	Exp	Pow	Val	Avg
Baseline [7]	0.151	0.122	0.031	0.207	0.128
Proposed Method	**0.379**	0.199	0.244	0.385	0.302
Nicolle et al. [18]*	0.354	**0.538**	**0.365**	**0.432**	**0.422**
Ozkan et al. [19]	0.117	0.076	0.062	0.200	0.114
Savran et al. [21]	0.306	0.215	0.242	0.370	0.283
Yang and Bhanu [24]	0.173	0.099	0.164	0.198	0.159

Video-only Testing Set					
Method	Aro	Exp	Pow	Val	Avg
Baseline [7]	0.077	0.128	0.030	0.134	0.093
Proposed Method	**0.302**	**0.244**	**0.199**	**0.279**	**0.252**
Nicolle et al. [18]**	-	-	-	-	-
Ozkan et al. [19]**	-	-	-	-	-
Savran et al. [21]	0.251	0.153	0.099	0.210	0.178
Yang and Bhanu [24]	0.190	0.105	0.142	0.177	0.154

*Best performing video feature.
**Video-only testing set not reported.

given in Section 4.5.1. Results comparing temporal feature methods on MMI-DB are given in Section 4.5.2. Testing results on AVEC 2011 and AVEC 2012 are given in Section 4.5.3. Testing results on CK are given in Section 4.5.4. A discussion on memory cost and visual examples of the proposed downsampling method are given in Section 4.5.5.

4.5.1 Selection of registration method and fusion rule

The selection of the best performing combination of registration method, and fusion rule is made with the development set on AVEC 2011. This experiment also tests the performance gain when using the proposed method versus a uniform sampling rate. The results for different registration techniques, sampling methods, and rules are given in Table 5. The methods are ranked in descending order of average performance across all four classes. Under sampling method, Uniform indicates that a uniform number of frames were selected for each segment, Proposed indicates that the proposed method was used. RST indicates that a similarity transform was used with eye points as control points. Sum refers to the sum rule; Product, product rule; Min, min rule; Max, max rule; Mode, the mode rule; and no fusion, the labels are assigned without any fusion. HMM indicates hidden Markov model fusion detailed in [47].

The best performer (Proposed + AIR + Sum) improves classification rate by 5.4% versus Uniform 1 + AIR + No fusion. This is the combination that is used in the following experiments, except for AVEC 2011 testing results, which are the original, official entry results of the challenge that used the Max rule. The combinations can be grouped into three categories: (1) dynamic downsampling with avatar image registration, (2) dynamic downsampling with similarity transform based registration, and (3) uniform downsampling with avatar image registration. It is clear that methods with the proposed dynamic sampling rate (groups 1 and 2) are better than methods that sample uniformly (group 3). While the two best performers use AIR registration, the difference between avatar image registration (group 1) and similarity transform registration (group 2) is not as clear. Replacing avatar image registration with similarity registration does not cause a significant drop in performance. Proposed + AIR + Sum and Proposed + RST + Sum have a difference of 1.4% on the average. For AVEC 2011, we conclude that intelligent selection of frames is a greater contributor to classification rate than a better registration algorithm.

4.5.2 Evaluation of temporal feature

We evaluate the use of optical flow as a temporal feature versus SIFT flow and frame differencing with MMI-DB empirically in Table 6. Weighted and unweighted accuracies are given in Table 7. When using a different temporal feature, ΔI_n is replaced by the new method (frame differencing or dense SIFT), the L_2-norm of the difference between frames n and $n-1$ is still used. For uniform sampling of 1 frame, the frame at the apex is the only frame used. Yang and Bhanu [24] is the worst performer because it uses all the frames, including the frames furthest away from

TABLE 10: Apex label results compared to other methods for 14 AUSs on CK. Bold indicates best F_1 performance, underline indicates second best. Avg: Average of all AUs.

	Facial Action Unit														
Method	**1**	**2**	**4**	**5**	**6**	**7**	**9**	**10**	**12**	**15**	**20**	**24**	**25**	**27**	**Avg.**
Proposed	85.3	93.0	**87.7**	69.6	**90.5**	62.4	68.5	43.5	76.9	**71.0**	74.0	**65.2**	93.6	84.2	**76.1**
Koelstra et al. [15]	86.8	90.0	73.1	**80.0**	80.0	46.8	**77.3**	48.3	83.7	70.3	**79.4**	63.2	**95.6**	87.5	75.9
Valstar et al. [46]	**87.6**	**94.0**	87.4	78.3	88.0	**76.9**	76.4	**50.0**	**92.1**	30.0	60.0	12.3	95.3	**89.3**	72.7
Yang and Bhanu [24]	82.0	92.1	82.0	58.6	84.9	52.5	68.4	34.8	68.2	66.7	65.7	51.1	85.6	67.2	68.6

the apex. Frame differencing is the fastest method for computing the temporal feature, but it has the worst performance among other temporal features. SIFT flow improves performance, but it is the slowest temporal feature optical flow has a better performance and speed. Retaining only 1 frame is worse than the proposed downsampling method. We conclude that, for MMI-DB, there are instances where retaining more than 1 frame can improve classification rate, if those frames are intelligently selected.

4.5.3 Results without apex labels

Results on the official AVEC 2011 testing and development sets are given in Table 8. The proposed method is compared to the two other entries that employed a dynamic sampling rate and it is always the best or second best performer for the development set. On the testing set, it improves weighted accuracy by 9.8%, and unweighted accuracy by 7.0% over the baseline approach. In [8], the method pays more attention when the predicted label changes, which assumes that the prediction is accurate, which is not always the case, especially for a difficult dataset such as AVEC 2011. We believe that the proposed method does well because it is the only downsampling method based on changes of visual information of the face.

Results on AVEC 2012 frame-level subchallenge are given in Table 9. Yang and Bhanu [24] is similar to the proposed approach but does not incorporate a downsampling and uses LPQ features. For the development set, Nicolle et al. [18] has the best performance, but they did not provide video-only testing results. They noted that the ground-truth labelers had a time delay when recording the label, and they incorporated metadata of who the user was speaking with, e.g. if the embodied agent speaking to them was belligerent. Though this improved performance, it is ad hoc in the sense that rater time delay may be specific to AVEC 2012, and meta-data about who the person is speaking to may not be available with other datasets.

4.5.4 Results with apex labels

The efficacy of the proposed method with apex labels on CK is given in Table 10. A comparison is made with other methods according to F_1 measure. For in-depth results see [35]. Yang and Bhanu [24] method does not take advantage of apex frame labeling. The proposed method takes advantage of apex labelling and it performs better. We performed best for 4 AUs. Valstar and Pantic [48] perform best for 6 AUs. However, the proposed method has a higher average F_1 measure among all the other works. Results for varying fusion rules, sampling methods and registration methods are given in [35]. The comparison to [24] demonstrates the importance of incorporating temporal information. Intuitively, assuming that each frame is equally discriminative, selecting as many frames as possible, such as in Yang and Bhanu [24], should increase the true positive rate by introducing more samples for the fusion. However, samples that are further away from the apex contain less relevant information of the expression being captured. Frames further away from the apex are close to neutral. They are not good examples of the expression being expressed, and they reduce accuracy. The proposed method sampled frames at the apex, and Koelstra et al. [15] modelled the temporal phases including the apex, this may explain the gap in performance.

TABLE 11: Summary of Frames Used for Each Dataset. Bold indicates least memory cost in terms of frames, underline indicates second best.

	AVEC 2011	**AVEC 2012**	**CK**	**MMI-DB**
# of Videos	74	95	488	222
# of Frames	1090476	1351129	8795	23466
Proposed	65871	76960	**1536**	**764**
Dahmane [8]	196051	239920	-	-
Savran [21]	-	232600	-	-
Glodek [13]	**740**	**950**	4930	2220

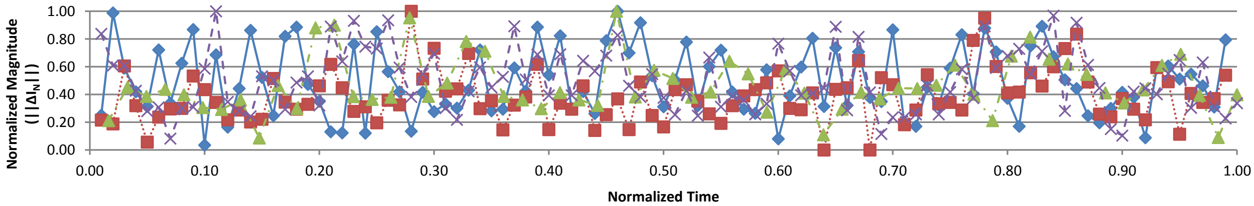
4.5.5 Memory cost savings and temporal feature results

In the following, we discuss the memory cost saving for each dataset, and show examples of the temporal feature. For AVEC 2011, the total number of frames for the development, training and testing (video subchallenge) partitions are $\{449074, 501277, 140125\}$, respectively. The proposed method downsampled the number of frames by a factor of 16.6, retaining $\{27412, 30076, 8383\}$ frames. For CK, the proposed method sampled 3.4 ± 2.2 frames. For MMI-DB the proposed method sampled 3.4 ± 1.5. A comparison of the number of frames reduced by the proposed method is given in Table 11.

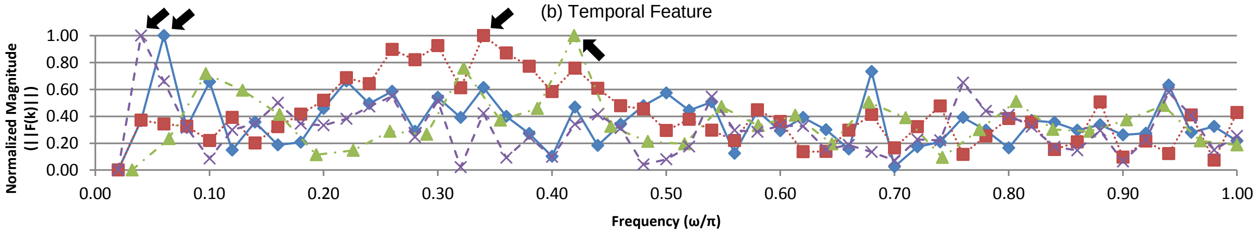
For a detailed explaination of the downsampling methods for related work, see Section 2.2. Because the method in [21] retains outliers based on the regression label, it can only be applied to continuous

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2014.2316151, IEEE Transactions on Affective Computing

(a) Examples of Continuous and Apex Labeled Video Segments



(b) Temporal Feature



(c) Fourier Transform of Temporal Feature

Fig. 6: (a) From top to bottom, a continuous video segment of a neutrally expressive person; a continuous video segment of an expressive person; an apex segment of a person who is expressive; an apex segment of a person who is less expressive. (b) The temporal feature of each of the examples, and (c) the discrete Fourier transform of the temporal feature. Note that both the continuous neutral and apex labeled less expressive examples have a low dominant frequency, whereas the other two expressive examples have a higher dominant frequency. Black arrow indicates dominant frequency.

label intensities, such as in AVEC 2012. The method would process each testing frame uniformly. In [8], for continuous data, we categorized the labels into 10 bins. This method is not applicable to apex labeled data, where the videos are segmented and have a single class label. In [13], frames are sampled uniformly. The method's memory cost is proportional to the number of videos, so the method does not reduce memory cost well for datasets with many videos, such as CK and MMI-DB. Though the method has the least number of frames for AVEC 2011 and AVEC 2012, it may sample the long videos too sparsely to precisely detect when emotion changes. The proposed method can be used to reduce the number of frames on all four datasets, both on continuous and discrete data, and on segmented and unsegmented data. It is the best or second best method for reducing memory cost on all four datasets.

A detailed example of two continuous video segments from AVEC 2012, and two apex labeled segments from MMI-DB is given in Figure 6. The magnitude has been normalized to provide a better understanding of the results. The time range has been normalized because MMI-DB segments are of different lengths. For the discrete Fourier transform, the frequency is normalized to $[0, 1]$. The first example in Figure 6(a) is of a person who does not use many expressions (Neutral). In this case the dominant frequency is at .06 cycles/frame, so only a few frames would be selected. The second row is of a person who is using many expressions and changing her pose (Expressive). Intuitively, many frames will be required to describe this segment, which is corroborated by the dominant frequency being at .34 cycles/frame. The third row is of a person who holds his expression for a long time at the apex (Apex Expressive). The dominant frequency is at .42 cycles/frame. In this example, there are 62 frames in the cycle, thus $.42 \times 62 \approx 26$ frames would be selected. It can be observed from the example frames that his expression is held at the apex for roughly half of the frames, corroborating keeping 26 of the 62 frames. The fourth row is of a person who weakly expresses his emotion (Apex Neutral). In this case, the dominant frequency is .04 cycles/frame, so very few frames would be selected.

5 Conclusions

In this paper, vision and attention theory was employed to temporally downsample the number of frames for video-based emotion and expression recognition. It was found that a uniform frame rate decreases performance and can unnecessarily increase memory cost for high frame rates. With the proposed method, AVEC 2011 is downsampled by a factor of 16.6 and weighted accuracy is improved over the baseline approach by 9.6% on the testing set. AVEC 2012 is downsampled by a factor of 17.6 and correlation is improved over the baseline by .159 on the testing set. CK is downsampled by a factor of 5.72 and the F_1 measure is improved by 0.3. MMI-DB dataset is downsampled by a factor of 30.1 respectively and weighted accuracy is increased over [24] by 8.4% for all sessions. Unlike previous works, we reported results on all four datasets.

The conventional process of using a short duration of frames centered at the apex was corroborated with the proposed sampling method, and extended to allow for an increase in duration when appropriate. It was found that top methods from previous challenges [24] did not generalize to continuous data sets. In that challenge, registration was found to be a significant contributor to performance, whereas, in the AVEC datasets, we have found that registration does not significantly contribute to performance. Previous datasets were segmented to the time points of most significance, and we posit that, for continuous datasets, a method must be critical in its selection of frames. A limitation of the current work is that the frames are processed in evenly sized segments, which may cause a boundary effect if an unlabeled apex is close to the segmentation boundary. However, this can be addressed by using overlapped boundary segments.

References

[1] G. McKeown, M. Valstar, R. Cowie, and M. Pantic, "The SEMAINE corpus of emotionally coloured character interactions," in *IEEE Conf. on Multimedia and Expo*, 2010.

[2] J. Yu and B. Bhanu, "Evolutionary feature synthesis for facial expression recognition," *Pattern Recognition Letters*, vol. 27, no. 11, pp. 1289–1298, 2006.

[3] P. Lucey, J. Cohn, T. Kanade, J. Saragih, and Z. Ambadar, "The extended Cohn-Kanade dataset (CK+): a complete dataset for action unit and emotion-specified expression," in *IEEE Conf. CVPR*, 2010.

[4] M. Valstar and M. Pantic, "Induced disgust, happiness and surprise: an addition to the MMI facial expression database," in *Corpora for Research on Emotion and Affect*, 2010.

[5] M. Valstar, M. Mehu, J. Bihan, M. Pantic, and K. Scherer, "The first facial expression recognition and analysis challenge," *IEEE Trans. SMC B*, vol. 42, no. 4, pp. 966 – 979, 2011.

[6] B. Schuller, M. Valstar, F. Eyben, R. Cowie, and M. Pantic, "AVEC 2011 the first international audio/visual emotion challenge," in *Affective Computing and Intelligent Interaction*. Springer Berlin / Heidelberg, 2011.

[7] B. Schuller, M. F. Valstar, F. Eyben, R. Cowie, and M. Pantic, "AVEC 2012 the continuous audio/visual emotion challenge," in *ACM Int'l. Conf. Multimodal Interaction*, 2012.

[8] M. Dahmane and J. Meunier, "Continuous emotion recognition using Gabor energy filters," in *Affective Computing and Intelligent Interaction*, ser. Lecture Notes in Computer Science. Springer Berlin / Heidelberg, 2011.

[9] Y. Zhu, F. De la Torre, J. F. Cohn, and Y. Zhang, "Dynamic cascades with bidirectional bootstrapping for action unit detection in spontaneous facial behavior," *IEEE Trans. Affective Computing*, vol. 2, no. 2, pp. 79–91, 2011.

[10] W. Tingfan, M. Bartlett, and J. Movellan, "Facial expression recognition using Gabor motion energy filters," in *IEEE Conf. CVPR*, 2010.

[11] P. Lucey, S. Lucey, and J. F. Cohn, "Registration invariant representations for expression detection," in *IEEE Conf. Digital Image Computing: Techniques and Applications*, 2010.

[12] D. Vukadinovic and M. Pantic, "Fully automatic facial feature point detection using gabor feature based boosted classifiers," in *IEEE Int'l. Conf. Systems, Man and Cybernetics*, 2005.

[13] M. Glodek, S. Tschenchne, G. Layher, M. Schels, T. Brosch, S. Scherer, M. Kachele, M. Schmidt, H. Neumann, G. Palm, and F. Schwenker, "Multiple classifier systems for the classification of audio-visual emotional states," in *Affective Computing and Intelligent Interaction*, ser. Lecture Notes in Computer Science. Springer Berlin / Heidelberg, 2011.

[14] B. Jiang, M. Valstar, B. Martinez, and M. Pantic, "A dynamic appearance descriptor approach to facial actions temporal modeling," *IEEE Trans. SMC B*, vol. PP, no. 99, p. 1, 2013.

[15] S. Koelstra, M. Pantic, and I. Patras, "A dynamic texture-based approach to recognition of facial actions and their temporal models," *IEEE Trans. PAMI*, vol. 32, no. 11, pp. 1940–1954, 2010.

[16] L. Maaten, "Audio-visual emotion challenge 2012: a simple approach," in *ACM Int'l. Conf. Multimodal Interaction*, 2012.

[17] H. Meng and N. Bianchi-Berthouze, "Naturalistic affective expression classification by a multi-stage approach based on hidden markov models," in *Affective Computing and Intelligent Interaction*. Springer Berlin / Heidelberg, 2011.

[18] J. Nicolle, V. Rapp, K. Bailly, L. Prevost, and M. Chetouani, "Robust continuous prediction of human emotions using multiscale dynamic cues," in *ACM Int'l. Conf. Multimodal Interaction*, 2012.

[19] D. Ozkan, S. Scherer, and L. Morency, "Step-wise emotion recognition using concatenated-HMM," in *ACM Int'l. Conf. Multimodal Interaction*, 2012.

[20] G. A. Ramirez, T. Baltrusaitis, and L. Morency, "Modeling latent discriminitive dynamic of multi-dimensional affective signals," in *Affective Computing and Intelligent Interaction*. Springer Berlin / Heidelberg, 2011.

[21] A. Savran, H. Cao, M. Shah, A. Nenkova, and R. Verma, "Combining video, audio and lexical indicators of affect in spontaneous conversation via particle filtering," in *ACM Int'l. Conf. Multimodal Interaction*, 2012.

[22] C. Soladie, H. Salam, C. Pelachaud, N. Stoiber, and R. Seguier, "A multimodal fuzzy inference system using a continuous facial expression representation for emotion detection," in *ACM Int'l. Conf. Multimodal Interaction*, 2012.

[23] M. Valstar, I. Patras, and M. Pantic, "Facial action unit detection using probabilistic actively learned support vector

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2014.2316151, IEEE Transactions on Affective Computing

machines on tracked facial point data," in *IEEE Conf. CVPR*, 2005.

[24] S. Yang and B. Bhanu, "Understanding discrete facial expressions in video using an emotion avatar image," *IEEE Trans. SMC B*, vol. 42, no. 4, pp. 920–992, 2012.

[25] M. S. Bartlett, G. Littlewort, M. G. Frank, C. Lainscsek, I. Fasel, and J. R. Movellan, "Automatic recognition of facial actions in spontaneous expressions," *J. Multimedia*, vol. 1, no. 6, pp. 22–35, 2006.

[26] T. Ojala, M. Pietikainen, and T. Maenpaa, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Trans. PAMI*, vol. 24, no. 7, pp. 971–987, 2002.

[27] L. P. Morency, A. Quanttoni, and T. Darrell, "Latent-dynamic discriminative models for continuous gesture recognition," in *IEEE Conf. CVPR*, 2007.

[28] J. Findlay and I. Gilchrist, *Active Vision: The Psychology of Looking and Seeing*. Oxford University Press, 2003.

[29] N. Ghosh and B. Bhanu, "A psychological adaptive model for video analysis," in *Int'l. Conf. Pattern Recognition*, 2006.

[30] R. Haber and M. Hershenson, *The Psychology of Visual Perception*. Rinehart and Winston Inc., 1973.

[31] P. Viola and M. Jones, "Robust real-time face detection," *Int'l. J. Computer Vision*, vol. 57, no. 2, pp. 137–154, 2004.

[32] C.-C. Chang and C.-J. Lin, "LIBSVM: a library for support vector machines," *ACM Trans. Intelligent Systems and Technology*, vol. 27, pp. 1–27, 2011.

[33] A. Jain, K. Nandakumar, and A. Rossb, "Score normalization in multimodal biometric systems," *Pattern Recognition*, vol. 38, no. 12, pp. 2270–2285, 2005.

[34] T. Gautama and M. A. V. Hulle, "A phase-based approach to the estimation of the optical flow field using spatial filtering," *IEEE Trans. Neural Nets*, vol. 13, no. 5, pp. 1127–1136, 2002.

[35] A. Cruz, B. Bhanu, and N. S. Thakoor, "Supplemental material for TAFFC submission 2013-03-0033," Online, 2013. [Online]. Available: URL to be Added

[36] T. Ahonen, A. Hadid, and M. Pietikainen, "Face description with local binary patterns: application to face recognition," *IEEE Trans. PAMI*, vol. 28, no. 12, pp. 2037–2041, 2006.

[37] G. Zhao and M. Pietikainen, "Dynamic texture recognition using local binary patterns with an application to facial expressions," *IEEE Trans. PAMI*, vol. 29, no. 6, pp. 915–928, 2007.

[38] H. T. Lin, C. J. Lin, and R. C. Weng, "A note on Platt's probabilistic outputs for support vector machines," *Machine Learning*, vol. 5, pp. 975–1005, 2004.

[39] (2013, September) Steam hardware & software survey: September 2013. Website. Valve Corporation. [Online]. Available: http://store.steampowered.com/hwsurvey

[40] E. Douglas-Cowie, R. Cowie, C. Cox, N. Amier, and D. Heylen, "The sensitive artificial listener: an induction technique for generating emotionally coloured conversation," in *LREC Workshop on Corpora for Research on Emotion and Affect*, 2008.

[41] G. McKeown, "Chatting with a virtual agent: the SEMAINE project character spike," Website, February 2011, http://www.youtube.com/watch?v=6KZc6e_EuCg.

[42] P. Ekman and W. Friesen, *Facial Action Coding System: A Technique for the Measurement of Facial Movement*. Consulting Psychologists Press, 1978.

[43] J. R. J. Fontaine, K. R. Scherer, E. B. Roesch, and P. C. Ellsworth, "The world of emotions is not two dimensional," *Psychological Science*, vol. 18, no. 12, pp. 1050–1057, 2007.

[44] P. Ekman and W. V. Friesen, "The repertoire of nonverbal behavior: categories, origins, usage, and coding," *Semiotica*, vol. 1, pp. 49–98, 1969.

[45] P. Ekman. (2013, December) What are micro expressions? Paul Ekman Group LLC. [Online]. Available: http://www.paulekman.com/me-historymore/

[46] M. F. Valstar and M. Pantic, "Fully automatic recognition of the temporal phases of facial actions," *IEEE Trans. SMC B*, vol. 42, no. 1, pp. 1 – 3, 2012.

[47] A. Cruz., B. Bhanu, and N. Thakoor, "Facial emotion recognition in continuous video," in *Int'l. Conf. Pattern Recognition*, 2012.

[48] M. Valstar and M. Pantic, "Fully automatic facial action unit detection and temporal analysis," in *IEEE Conf. CVPR*, 2006.

Albert C. Cruz received the BS degree in electrical engineering from the University of California, Riverside, CA, USA in 2008. He is currently a PhD student in Electrical Engineering at the Center for Research in Intelligent Systems at the University of California, Riverside, CA, USA. His research interests include image processing, computer vision, pattern recognition and machine learning. His current research focuses on biologically-inspired algorithms for facial emotion and expression recognition.

Bir Bhanu received the SM and EE degrees in electrical engineering and computer science from the Massachusetts Institute of Technology, Cambridge, the PhD degree in electrical engineering from the Image Processing Institute, University of Southern California, and the MBA degree from the University of California, Irvine. He is a distinguished professor of electrical engineering and a cooperative professor of computer science and engineering, mechanical engineering, and bioengineering, and the director of the Center for Research in Intelligent Systems and the Visualization and Intelligent Systems Laboratory at the University of California, Riverside. He is also the director of NSF IGERT on Video Bioinformatics at UCR. His research interests include computer vision, pattern recognition and data mining, machine learning, artificial intelligence, image processing, image and video database, graphics and visualization, robotics, human-computer interactions, biological, medical, military, and intelligence applications. He has been the principal investigator of various programs for the NSF, DARPA, NASA, AFOSR, ARO, ONR, and other agencies and industries. He is a fellow of the IEEE, AAAS, IAPR, and SPIE.

Ninad S. Thakoor received the B.E. degree in electronics and telecommunication engineering from the University of Mumbai, Mumbai, India, in 2001 and the MS and PhD degrees in electrical engineering from the University of Texas at Arlington, TX, USA, in 2004 and 2009, respectively. He is a Postdoctoral Researcher with the Center for Research in Intelligent Systems, University of California, Riverside, CA, USA. His research interests include vehicle recognition, stereo disparity segmentation, and structure-and-motion segmentation.