ADVANTAGES OF DYNAMIC ANALYSIS IN HOG-PCA FEATURE SPACE FOR VIDEO MOVING OBJECT CLASSIFICATION

Miriam M. López *Lucio Marcenaro* *Carlo S. Regazzoni*

DITEN, University of Genova
Via Opera Pia 11A, 16145 Genoa - Italy
{miriam.lopez, mlucio, carlo}@ginevra.dibe.unige.it

ABSTRACT

Classification of moving objects for video surveillance applications still remains a challenging problem due to the video inherently changing conditions such as lighting or resolution. This paper proposes a new approach for vehicle/pedestrian object classification based on the learning of a static kNN classifier, a dynamic Hidden Markov Model (HMM)-based classifier, and the definition of a fusion rule that combines the two outputs. The main novelty consists in the study of the dynamic aspects of the moving objects by analysing the trajectories of the features followed in the HOG-PCA feature space, instead of the classical trajectory study based on the frame coordinates. The complete hybrid system was tested on the VIRAT database and worked in real time, yielding up to 100% peak accuracy rate in the tested video sequences.

***Index Terms*—** Moving object classification, HOG, PCA, HMM, hybrid classifier.

1. INTRODUCTION

Classification of moving objects is a fundamental step that still represents an active research area. Along with detection and tracking, object classification can lead to a more complete understanding of the scene, thus allowing the enhancement of the automation in video applications such as public security surveillance. In recent years, a great effort has been done by many researchers addressing this task [1, 2]. However, real-world applications present particular challenges in which high accuracy, flexibility to continuously changing environmental conditions and real-time operation are required.

Systems aimed at automatic moving object classification developed so far can roughly be categorized into two groups. The first group includes systems that perform object detection without prior segmentation, in which one specific target object, such as vehicles or faces [3], is required to be detected in the scene. Strictly speaking, these are not classification systems but detectors, as they cannot identify more than one learned pattern. Yet efficient in terms of computation time, these methods present some limitations specially when the target pattern varies. In this line of research, a special focus has been done on pedestrian detection [4, 5, 6]. The second group consists of methods that first identify moving objects in the scene and subsequently categorize them into one of the predefined classes [7]. These classification approaches are usually part of more complex systems in which a tracker is also implemented, for instance, based on background subtraction. A special attention has been paid to distinguishing pedestrians from other moving objects – usually vehicles – in traffic scenes [2].

A wide array of possibilities has been proposed in the literature as feature extraction techniques for object classification purposes, ranging from features based on raw appearance [8], texture [9], shape [10] or a combination of some of them [11]. In all these cited works, features are proposed as *static* features, that is, no temporal correlation is exploited by these features, and the extraction process has to be performed on the current frame independently of the results obtained in the previous frames.

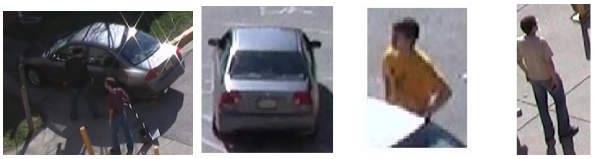
In the last few years, there has been an increasing awareness of the necessity of using *dynamic* models that somehow integrate classification within a discrete Markovian process, as motivated by the discipline of cognitive dynamic systems. The dynamic methods used for modelling the video scene proposed in the literature are mainly based on the spatio-temporal information given by frame coordinates of the object. For instance, motion can be useful in some scenarios for distinguishing between people and vehicles [7]. However, it would not allow the identification of more subcategories such as cars, trucks or vans. Learning trajectories followed by different objects in the scene has also been proposed for object classification [12]. These approaches usually make use of the object position coordinates – and sometimes the first and second derivatives, to make up a flow vector that contains spatio-temporal information of each moving object [13]. By collecting motion data over sufficiently long time slots, trajectories can be learnt by means of clustering methods [14] or Hidden Markov Models (HMM) [15], which have been previously used for diverse applications such as interaction analysis [16] or identification of abnormal situations [17]. It has been shown that classification benefits from the time regularization effects on the position information in such cases. However, the definition of 'object state' can be extended to include other object properties in addition to dynamic position evolutions (i.e. motion trajectories).

One important limitation of the mentioned dynamic systems concerns the assumption that objects move along fixed trajectories throughout the scene [13]. In this work, we tackle the pedestrian/vehicle classification problem in video scenarios in which this assumption does not hold. We propose a hybrid system that consists of two main classification modules, one static and one dynamic, that run concurrently and complement each other. The main contribution lies in the extension of the definition of object state to a time variant appearance property, in order to evaluate to what extent time regularization on other object state subspaces can provide benefits to classification. This is achieved by analysing the trajectories in a content-based feature space, contrary to the use of frame coordinates or motion information as proposed in the literature so far. By defining a fusion decision rule, the static and dynamic modules are efficiently combined and a final label is emitted for the detected moving object.

The remainder of this paper is organized as follows: in Section 2 the database and the preprocessing steps are described, as well as the

Clip	Class	# Training objects	# of training samples	# of testing samples
1	Pedestrian	3	145	2176
	Car	2	220	2120
2	Pedestrian	7	342	7828
	Car	1	51	1500
3	Pedestrian	3	145	7828
	Car	2	220	1500

Table 1: Number of selected objects, training and testing samples for each experiment.



(a) (b) (c) (d)

Fig. 1: (a) Overlap among objects. (b)-(d) Three real image patches used in this work.

feature extraction approach used. In Section 3, the proposed hybrid classification system is explained in detail. Experimental results are shown and discussed in Section 4. Finally, in Section 5 possible future work lines are mentioned and conclusions are drawn.

2. IMAGE PREPROCESSING AND FEATURE SPACE

The video sequences used in this work are extracted from the VIRAT database [18], which has been recently used for object classification purposes [2]. One of the main reasons for using this database is that it presents real scenarios in which cars and pedestrians moves freely through the same areas and at comparable speeds, and therefore, frame coordinates and motion information are no longer discriminant features, as highlighted in the introduction. Three video sequences (clips from now on) are selected from the VIRAT database for testing the proposed approach. For clips 1 and 2, videos used in the training and testing steps belong to the same video sequence, yet different for each video clip. In video clip 3, the training step is performed using samples of one video sequence and the testing is performed on a different sequence. In any case, a frame that is used for the training step is never used for the test. The number of training samples has been limited on purpose to a reduced number in order to present a realistic approach. Details on the number of training and testing objects are shown in Table 1.

2.1. Image preprocessing

Given the large amount of works present in the literature concerning multi-object tracking methods [19], we assume that a region with a rectangular shape containing a detected object of interest is provided by the tracker. In order to build as view-independent as possible system, a 'spatial normalization' method in terms of homogenization of the image patch dimensions is first applied, leading to a fixed-dimension square containing the object of interest. This step is done with the purpose of a subsequent application of a multivariate feature extraction method, allowing the extraction of the same number of features for all objects regardless their shape or dimensions. The scaled patch dimensions were set to $hight = width = 128$, which experimentally were found to be a suitable value.

2.2. Input space features

In this work we exploit the idea that in a real video sequence, the *content* of image patches corresponding to the same moving object is expected to change smoothly along time. One of the most well-known descriptors capable to extract image content related information is the Histogram of Oriented Gradients (HOG), which was first proposed by Dalal and Triggs [5] for people detection in video. Specifically, HOG captures edge or gradient structure that is very characteristic of the local shape and has been widely used for feature extraction [20]. HOG has been traditionally used in object detection tasks, which means that the whole frame is examined while looking for the learned target pattern. In this work, as for the other tested features, HOG is applied as an extractor of significant values from the image patch that will be used as input for the classifier. That means that, contrary to how it was originally presented, the HOG operation is applied to image patches of both classes, pedestrians and vehicles, and not to a single target pattern.

2.3. PCA feature space

The application of the HOG descriptor to the image patches yields as a result a set of features that can be rearranged into a vector of size n. The dimension of this vector can be efficiently reduced to a dimension $m << n$ by means of Principal Component Analysis (PCA) [21], which consists of a linear transformation of the original data.

3. THE PROPOSED METHOD

We propose to study the underlying movement patterns of the image patch content change in the feature space given by the PCA coefficients of the HOG features. Figure 2 depicts this idea. The figure shows the trajectories in the HOG-PCA feature space followed by features corresponding to a pedestrian walking and a car moving in the scene. For the car (red asterisks), features move in a smooth and more linearly way because the patch content barely changes if the camera viewpoint remains the same, except for the background pixels, whereas for the pedestrian, features lie in a closer area due to the cyclic nature of the pedestrian walking movement.

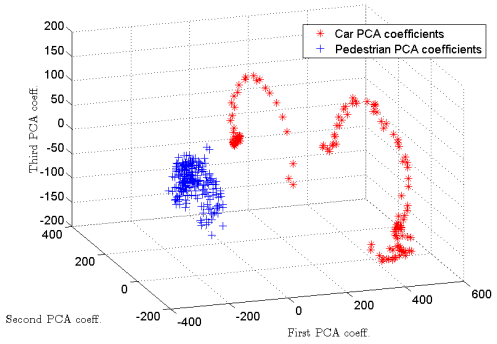


Fig. 2: Pedestrian and car feature trajectories in the HOG-PCA space.

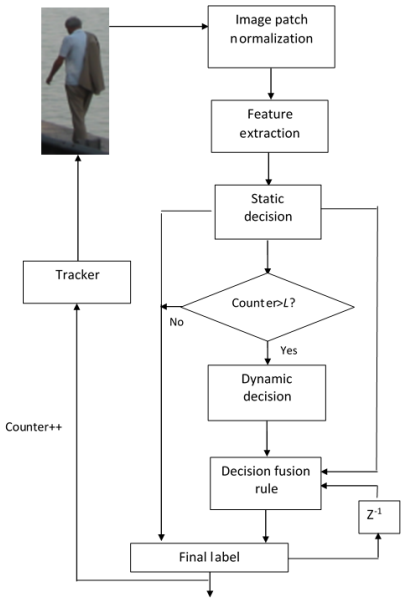


Fig. 3: General scheme of the proposed hybrid system.

The scheme of the proposed system is shown in Figure 3. When a new moving object is detected in the scene, HOG features are extracted and compressed using the PCA projection matrix computed in the training step. The resulting HOG-PCA coefficients are used to feed a static classifier, which immediately emits a label . The static classifier decision is contrasted with the output of an HMM-based dynamic module, which runs simultaneously and analyses the trajectories of these HOG-PCA coefficients. The only restriction for the dynamic module is to wait for a number L of frames, in order to give enough time to understand the dynamics of object features. Details on each classification module are given below.

3.1. Static classification

The kNN classifier is chosen for its simplicity and fast output computation, which is essential for real time applications. In order to understand which value of k best performs for the vehicle/pedestrian classification problem, different values of $k = 3, 5, 7, ..., 21$ have been tested in the experiments.

3.2. Dynamic model

The HMM is a generative model that has been widely used for representing the dynamics of the scene for video applications [15]. An HMM can be compactly represented as:

$$\lambda = (A, B, \pi) \tag{1}$$

and is designed to have Q states. The transition matrix $A = \{a_{ij}\}, i, j = 1, ..., Q$ explains the probability of transition from state $i \rightarrow j$. Each state has a probability distribution function over the possible output observation, denoted by the elements of $B = \{b_j\}$. Finally, π_i denotes the probability for the system to start in state i. As the possible values in the feature space are not restricted to a limited set of discrete values, we propose the use of continuous density HMMs in which each emitted observation is generated according to a PDF dependent on the state at each time instant. This PDF is normally modeled as a mixture of M Gaussians. Thus, the state-conditional observation PDF is given by:

$$b_j(O_t) = \frac{1}{(2\pi)^{m/2}|\Sigma_j|^{1/2}} \exp\left\{-\frac{1}{2}(O-\mu_j)^T \Sigma_j^{-1}(O-\mu_j)\right\} \tag{2}$$

where the Gaussian is a multivariate normal distribution of the same dimensionality m as the number of PCA coefficients representing the object. The parameters of the HMM are initialized to random values and the Baum-Welch algorithm is used for estimation using the forward-backward procedure [15]. In our experiments, we set $M = 1$ because we assume just a single typical behaviour for each object, and the number of states is set to $Q = 2$. A higher number of states would increase the computational time remarkably, making the system unfeasible for real-time applications.

For classification purposes, one HMM must be trained for each class. The learning problem basically consists of finding the parameters of a HMM model $\lambda_k = (A, B, \pi)$ given a sequence of observations $O = O_1, ..., O_T$. The target of the learning problem is to maximize $P(O|\lambda)$, which can be solved using the Baum-Welch algorithm [15]. Once the HMMs for all classes have been trained, the classification of new objects can be performed by computing the likelihood of each HMM for describing the test feature trajectory O_{test} in the HOG-PCA feature space. The assigned class is the one associated to the HMM that maximizes the likelihood given the test sequence O_{test}. Regarding the length of the observation sequence L, different values ranging from $L = 2, ..., 5$ have been tested, as we expect to find the movement patterns in a short time slot, specially those regarding pedestrians.

3.3. Decision fusion rule

During the first L frames in which a new object comes into view, the static label is assigned to the object as it is the only label available. After the L-th frame, a decision rule that considers the static and the dynamic outputs is applied. If at time instant k both classifiers agree on the label of the object as y_k, and it also matches the one emitted for this object in the previous time instant y_{k-1}, then that label will be assigned to the object at time k as well. If the classifiers do not agree, or they do but there has been a change with respect to the previous label assigned to the object, a conflict occurs. In such case, the conflict is resolved by picking the classifier with highest *self-confidence* at time instant k, which is computed as the number times in the past b frames (time instants $k-1, k-2, ..., k-b$) that the emitted label was the same as at time k. In case of conflict and equal self-confidence of both classifiers, the final label is determined by a majority voting performed on the labels assigned by both classifiers in the past b frames. The buffer size is set to $b = 5$ in the experiments.

4. EXPERIMENTAL RESULTS

A collection of classification results is shown in Figure 4. In each graph, accuracy curves obtained by the static classifier (represented for comparison purposes as it represents the classical classification system) and by the proposed hybrid system are shown together. Notice the double x axis, representing each one an inherent parameter of the static and the hybrid classifiers, namely the number of neighbours k in black colour for the static classifier, and the sequence length L in blue colour for the hybrid system, respectively. The

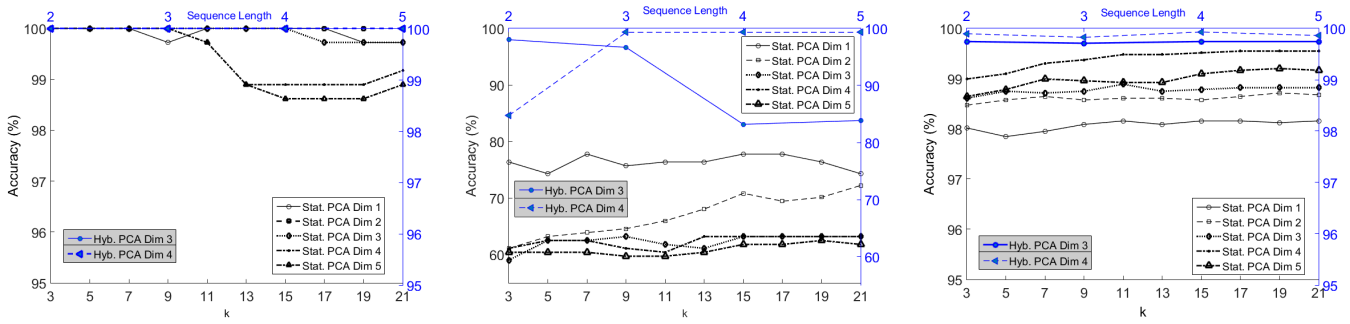


Fig. 4: Accuracy results given by the static classifier (black line) and the proposed hybrid classifier (blue line) for video clips 1 (left), 2 (central) and 3 (right).

curves are depicted in the same colour as their corresponding x axis, and accuracy rate is represented in the y axis for both classifiers.

The accuracy curves make evident that the hybrid system outperformed (or at least equalled) the static classifier performance in all experiments. Furthermore, there is a low dependency of the hybrid classifier on the whole system internal parameters, that is, its superiority happens for all values of nearest neighbours k, sequence lengths L, and for different PCA feature space dimensions, being $m = 3$ the number of PCA coefficients that yielded best results in more cases for all video clips, but not the only possible value that performed satisfactorily.

Overall, the hybrid system provided better results for clip 1 than for clip 2. This is a fair result given the characteristics of the training and testing number of objects used in each one: in clip 2 only one car and seven pedestrians appear during the training stage, whereas the training dataset for clip 1 is more balanced (see Table 1 for details). The introduction of the fusion rule allowed for the correction of isolated classification errors made by the single classifiers, reaching up to 100% accuracy for different parameters of the system tested on clip 1. For video clip 2, the improvement introduced by the hybrid system is highly noticeable with respect to the static classifier, increasing the accuracy rate from 77.78% (peak value obtained for $k = 7$) to 99,31% (value obtained for sequence lengths $L = 3, 4, 5$). The performance obtained on clip 3 proves that the whole system is robust to scene changes.

The computational cost of the complete hybrid system is assessed by measuring the time needed for different numbers of principal components (m) used for both the dynamic and static classifiers. Results are shown in Table 2. Videos of VIRAT database are recorded at frame rates ranging $25 - 30Hz$, which means that methods requiring times over $0.034 - 0.040s$ are not suitable for real time execution. However, it is worth highlighting that for the dynamic part of the system, introducing a frame step so that not all the frames are processed but one every $3 - 5$ frames is not only acceptable, but even recommendable, as skipping some frames allows the learning of the dynamics better than with a frame-by-frame approach, in which content changes are hardly noticeable and therefore, longer sequence lengths are needed for the learning step. As expected, there is an increasing demand of computational time when the number of objects to be processed simultaneously in one frame increases. PCA-HOG features showed real time performance up to 5 objects appearing simultaneously in the scene. All the reported time measurements correspond to the average time computed over all the frames showing the same number of objects.

Finally, we would like to point out that, even though finding the optimal features for the vehicle/pedestrian classification task is out of the scope of this work, other content-based local features such as 2D wavelet coefficients and Canny and Sobel detectors were also tested in the same way as HOG descriptor was. Complete set of results is not shown due to space limitations but in all cases, the hybrid system outperformed the static one, and HOG descriptor proved to be the most suitable one among all the mentioned feature extraction methods tested for this specific classification task.

Number of PCs	Number of objects in the scene						
	1	2	3	4	5	6	7
$m = 3$	0.0087	0.0150	0.0200	0.0253	0.0372	0.0423	0.0499
$m = 4$	0.0086	0.0152	0.0215	0.0262	0.0406	0.0465	0.0571
$m = 5$	0.0088	0.0148	0.0201	0.0257	0.0411	0.0481	0.0527

Table 2: Average computational times (secs) for the hybrid system for different number of objects appearing simultaneously in the scene.

5. CONCLUSIONS AND FUTURE WORK

In this work, a system for vehicle/pedestrian classification is proposed, in which the strengths of static and dynamic classifiers are joint to generate a more accurate hybrid classifier for video applications. The study of the dynamic aspects of the moving objects is introduced in a novel way, in which a HMM learns the underlying trajectory patterns followed by content-based HOG-PCA features, instead of by trajectories of objects themselves in the scene. The decisions cast by both classifiers are combined by means of a fusion decision rule, which shows a considerable improvement in terms of classification accuracy results.

An extension of the proposed method for a multi-object framework may be easily achieved, as both the kNN and the HMM-based model can be trained for a multi-class task. Computational time should be then considered carefully, specially for the dynamic module, which might make the whole system become unfeasible for real-time applications.

6. REFERENCES

[1] Yi Liu and Yuan F. Zheng, "Video Object Segmentation and Tracking Using ψ-Learning Classification," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 15, no. 7, pp. 885–899, 2005.

[2] Mohamed Elhoseiny, Amr Bakry, and Ahmed Elgammal, "Multiclass Object Classification in Video Surveillance Sys-

tems - Experimental Study," *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 788–793, June 2013.

[3] Paul Viola and Michael Jones, "Rapid object detection using a boosted cascade of simple features," *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, vol. 1, pp. I–511–I–518, 2001.

[4] Paul Viola, Michael Jones, and Daniel Snow, "Detecting pedestrians using patterns of motion and appearance," in *Ninth IEEE International Conference on Computer Vision (ICCV)*, 2003, vol. 2, pp. 734–741.

[5] Navneet Dalal and Bill Triggs, "Histograms of Oriented Gradients for Human Detection," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, 2005, pp. 886–893.

[6] Piotr Dollár, Christian Wojek, Bernt Schiele, and Pietro Perona, "Pedestrian Detection: An Evaluation of the State of the Art," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 4, pp. 267–272, 2012.

[7] Paulo Vinicius and Koerich Borges, "Pedestrian Detection Based on Blob Motion Statistics," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 23, no. 2, pp. 224–235, 2013.

[8] Y. Bogomolov, G. Dror, S. Lapchev, E. Rivlin, and M. Rudzsky, "Classification of Moving Targets Based on Motion and Appearance," *Procedings of the British Machine Vision Conference 2003*, pp. 44.1–44.10, 2003.

[9] Matti Pietikinen Timo Ojala and David Harwood, "A comparative study of texture measures with classification based on featured distributions," *Pattern Recognition*, vol. 29, pp. 51–59, 1996.

[10] Damian Ellwart and Andrzej Czyzewski, "Viewpoint independent shape-based object classification for video surveillance," in *12th International Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS)*, 2011.

[11] Zhaoxiang Zhang, Yinghao Cai, Kaiqi Huang, and Tieniu Tan, "Real-Time Moving Object Classification with Automatic Scene Division," in *IEEE International Conference on Image Processing (ICIP)*, 2007, vol. 5.

[12] Faisal I. Bashir, Ashfaq A. Khokhar, and Dan Schonfeld, "Object Trajectory-Based Activity Classification and Recognition Using Hidden Markov Models," *IEEE transactions on Image Processing*, vol. 16, no. 7, pp. 1912–1919, 2007.

[13] Brendan T. Morris and Mohan M. Trivedi, "Learning and Classification of Trajectories in Dynamic Scenes: A General Framework for Live Video Analysis," in *IEEE Fifth International Conference on Advanced Video and Signal Based Surveillance*, 2008, pp. 154–161.

[14] Morris Brendan and Mohan Trivedi, "Learning Trajectory Patterns by Clustering: Experimental Studies and Comparative Evaluation," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009, pp. 312–319.

[15] Lawrence R. Rabiner, "A tutorial on HMM and selected applications in speech recognition," *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257–286, 1989.

[16] Alessio Dore and Carlo S. Regazoni, "Interaction Analysis with a Bayesian Trajectory Model," *IEEE Intelligent Systems*, vol. 25, no. 3, pp. 32–40, 2010.

[17] A. Cavallaro F. Daniyal, "Abnormal motion detection in crowded scenes using local spatio-temporal analysis," in *Proceedings of the 36th IEEE International Conference on Acoustic, Speech and Signal Processing (ICASSP)*, 2011, vol. 3, pp. 1944–1949.

[18] Sangmin Oh, Anthony Hoogs, Amitha Perera, Naresh Cuntoor, C.-C. Chen, Jong Taek Lee, Saurajit Mukherjee, J. K. Aggarwal, Hyungtae Lee, Larry Davis, Eran Swears, Xioyang Wang, Qiang Ji, Kishore Reddy, Mubarak Shah, Carl Vondrick, Hamed Pirsiavash, Deva Ramanan, Jenny Yuen, Antonio Torralba, Bi Song, Anesco Fong, Amit Roy-Chowdhury, and Mita Desai, "A large-scale benchmark dataset for event recognition in surveillance video," in *IEEE Conference on Comptuer Vision and Pattern Recognition (CVPR)*, 2011, pp. 3153–3160.

[19] Luigi Di Stefano Samuele Salti, Andrea Cavallaro, "Adaptive Appearance Modeling for Video Tracking: Survey and Evaluation," *IEEE Transactions on Image Processing*, vol. 21, no. 10, pp. 4334–4348, 2012.

[20] Xiaodan Zhuang P. Natarajan Huaigu Cao A. Jain, Xujun Peng, "Text detection and recognition in natural scenes and consumer videos," in *Proceedings of the 39th IEEE International Conference on Acoustic, Speech and Signal Processing (ICASSP)*, 2014, pp. 1245–1249.

[21] I. T. Jolliffe, *Principal Component Analysis*, Springer-Verlag, 1986.