Anthropocentric Video Analysis For Film And Games Postproduction

Ioannis Pitas, Nikolaos Nikolaidis

Abstract: The interest of the scientific community for anthropocentric (human-centered) video analysis stems from the fact that the extracted information (e.g. human presence, identity, body posture, emotional status, body parts movements, activities) can be utilised in various important applications. One such application domain is film and games postproduction, where the anthropocentric video analysis results can be used in various tasks, such as audiovisual material indexing and retrieval or automatic semantic annotation. In this paper, we shall review recent research results in a number of diverse areas, such as face/person detection, human activity recognition and face/facial expression recognition, either from a single or multiview visual (image/video) sources.

Key words: *human-centered computing, human-centered interfaces, computer games, film postproduction, video analysis.*

INTRODUCTION

During the last two decades, we have witnessed an increasing research interest towards the so-called *anthropocentric (human-centered) video analysis*, namely, algorithms that aim to extract, describe and organize information regarding the basic element of most videos: humans. This diverse group of algorithms processes videos from various sources and extracts a wealth of useful information related to the state (presence, identity, body posture, emotional status, etc.) and state transitions (body parts movements, human activities, etc.) of individuals, to interactions or communication modes between two or more humans (dialogues, social signals, etc.) and to the physical characteristics of humans, such as 3D body models.

Since the anthropocentric video analysis covers so many diverse topics [1], we focus in this survey only to recent research results in four different areas, notably human detection (single view and multiview), human activity recognition and a subspace approach to face and facial expression recognition.

HUMAN DETECTION AND POSE ESTIMATION

The human detection task has been approached using two basic methods and their combination. The first basic method is template matching using artificially generated human body silhouettes [2,3] and the second one uses the Histogram of Oriented Gradients (HOG) features [4] as input to an SVM classifier. Their combination uses template matching to find a silhouette (that could correspond to a person) and then extracts the HOG features at relevant points that are distributed along the silhouette instead of using the HOG features of the entire region of interest.

The template matching approach [2,3] employs a shape similarity measure (a variation of the Modified Hausdorff Distance) to assess how closely a shape template (in this case a human silhouette) matches the edges of the input image. In order to speed up the matching process, a binary search tree data structure (*shape tree*) is used, since the exhaustively matching to every template would be too slow. This data structure stores every template in its leaf nodes. First, we obtain an image (video frame) edge map, in order to detect human silhouettes therein. Then, we try to detect humans in various consecutively chosen image windows. When an input window of the images edges is to be tested, the edges of the window are used at the shape tree root and every subsequent internal node to determine towards which sub-tree the search will continue. When a shape tree leaf node is reached, this node's template is considered as a candidate for a possible human silhouette match. The search can, at user's discretion, backtrack and reverse a

decision along the path from the tree root to the leaf node, so that it can reach a different leaf node and, therefore, find another matching human silhouette candidate. This can be repeated several times (specified by the user), so that a list of matching human silhouette candidates is formed that contains only a small fraction of all available silhouette templates. This list is then exhaustively matched with the edges of the input window. If the similarity measure of a human silhouette candidate is sufficiently high, a positive human detection decision is taken. The human silhouette search is handled by first properly scaling the image (video frame), so that the height of the persons in the scaled image is comparable to the height of the silhouette templates. The search scans the edge image of the scaled input image and inputs the edges of the search window into the shape tree. After the search process is finished, candidate silhouettes that are close to each other (in position) are merged to the one having the highest similarity score. Example results are shown in Figure 1. As a side-result, the shape tree human detection method provides also human body posture estimation.

Figure 1: Detected human bodies.

The HOG human detector [4] builds a pyramid of images from the input image (video frame) at various scales. At each scale, the HOG features are computed by splitting the image into 8x8 non-overlapping cells. Inside each cell, the image gradient of each pixel (obtained from a Sobel mask) is added to one of nine possible orientation bins. Each scaled image is scanned using a sliding window approach. At each position of the search window, every 2x2 block of cells (these blocks can have overlapping cells) is concatenated into a feature vector. This feature vector is then classified by an appropriately trained SVM, as either representing the HOG of a human body or not. At the end of the search, results with significant location overlap are merged together.

The combination of the above two methods [5] has been applied to the task of detecting human heads, as a first step to a top-down body pose estimation system. Instead of full body silhouettes, artificially generated head and shoulders silhouettes are stored inside a shape tree. During the training process, the head shape tree is used to find the best template that matches the subject's head in the training image. The corresponding 2x2 HOG blocks at locations that coincide with the matched template are concatenated into a feature vector. In a similar way, negative examples are gathered by applying the same process to input images (video frames) that are known to contain no human heads. The gathered data are used to train an SVM classifier. During testing, we first use the tree to find a silhouette match, then concatenate the relevant HOG blocks into a feature vector, which is then used as input to the trained SVM. If the SVM verifies the presence of a human head, a positive detection decision is taken. Closely located detection results are merged to the head shape candidate with the highest similarity score.

MULTI-VIEW OBJECT, HUMAN BODY AND BODY PART DETECTION

Single view person detection tends to produce several false positive body detections. Therefore, if we have multi-view video data (as is the case in many film/game productions), it is advisable to use all views for person detection. In this section, we shall present a new method for multi-view human body (or body part, or general object) detection. The basic idea is to use a single view object detector in every view of a scene captured by multiple cameras and then combining the results using the 3D information of the scene/cameras. The method can improve the results of the single view object detector, while also localizing the object/human in the 3D space. This results in a robust way for rejecting the false detections, amending the missed detections and associating the results of the single view detector across views.

The successful object (or human body/body part) detection in videos has many applications that include tracking, object/human recognition, activity recognition, surveillance and robot vision. Numerous object or human body (or body part) detectors that operate on a single image or single view video have been developed in the past few years, having a fair success rate, as described in the previous section [2-5]. However, the human or object detection in a convergent multiple camera environment has scarcely been touched upon, although such algorithms could find important applications in stereoscopic cinema or TV production and postproduction, e.g. by providing useful information for person matting initialization or camera calibration (by excluding moving humans from the procedure). Obviously, the existence of multi-view information is expected to lead to improved detection results. However, most existing methods do not rely only on this information, in order to achieve a 3D object detection by matching detections across views. In [6], the authors make use of color histograms to associate data across different views. Human body part representations, such as upper and lower arm representations by line segments are used in [7] for matching such structures across views. In [8], the 3D human body detection is trivial, since the scene captured is a football field providing easy background extraction and a ground plane that limits the 3D search space. Calibration information is used in [9] to match the object detections across views, though the utilized ray intersection technique requires a very accurate single view detector.

We recently proposed a novel multi-view detection method [10] that utilizes a single view detector to locate objects/humans in each one of the multiple views of the same scene, obtained through calibrated cameras. The proposed method uses this information, along with the calibration information, in order to locate the object in the 3D space and also improve the detection results in each view, by eliminating false positive and false negative detections and associating detected objects, bodies or body parts (e.g. faces) across views. The basic idea behind the proposed method is the following. Let us assume that we have an object in a scene and a number of images of this scene, some of which depict the object in question. We assume that the images have been obtained by a set of convergent calibrated synchronized cameras. Then, for every image, there exists a projective mapping that relates the 3D coordinates of the object to the 2D coordinates of the object projection on each image plane. These mappings provide a unique way to fuse the 2D object location information in every image derived by the application of a certain 2D object detector. Thus, the effects of occlusion are minimized, 3D estimates of the object location are provided and the accuracy and robustness of the 2D and 3D object location estimation are improved. At least two or more cameras, calibrated with respect to a common coordinate system should be available. Special cases of objects can be human bodies or body parts (e.g. human face, head). The procedure followed by the proposed multi-view object and human body part detector can be summarized in the following steps:

2D object detection: Having a set of images depicting a scene from different views, we use a single view detector to locate objects, human bodies or body parts, resulting in correct or false detections.

Voting and 3D object detection: By back-projecting each detected object (human body or body part) view to the 3D space utilizing camera calibration information, we find the 3D regions, where the created projection volumes intersect each other. Using a voting approach, we find which of these regions collect enough votes and correspond to the scene entities (objects). Thus, each selected 3D region corresponds to a single entity. A set of detected instances of this entity on each view is also associated with the 3D region. This step allows us to reject false detections in the various views.

2D object view associations: In this step, information from the voting step is used to associate 2D object detections across views, so that all associated detections correspond to the same 3D entity and also to detect the entity projection in views that the single view 2D object detector failed to do so.

This method was used to implement a multi-view human head detector [10]. This detector utilizes a variation of the face detector proposed in [11], trained to detect both frontal and profile faces. The frontal/profile detector in [11] was improved in two ways. First the detector was applied in rotated versions of the input image in order to detect rotated faces (e.g. tilted heads) resulting in more true positive (but also some false positive) face detections. Then, a skin color detector was used in order to reject the false positive detections, as was done already in [12]. In more detail, the pixel color in the facial bounding boxes (BB) returned by the frontal/profile face detector [11] were checked against a range of skin-like colors. BBs that contained a small percentage of skin-like colored pixels were rejected. Finally, the multi-view detector was used to reject the false detections, amend the missed detections and associate the results of the single view detector across views. Results of the multi-view head detector are presented in Figure 2, where red denotes false (subsequently rejected) detections, blue denotes missed (subsequently rectified) detections, green denotes correct detections and numbers denote the associated face detections corresponding to the same detected 3D heads.

Figure 2: Multiview head detection in two different views.

HUMAN ACTIVITY RECOGNITION

Human activity recognition is an important task in the semantic video analysis that, typically, follows human body detection and body pose estimation. Our approach [13-15] in movement recognition exploits the binary masks resulting either from a background subtraction, or a chroma keying technique, or from human body posture estimation, as described in the first section of this paper. Such masks depict the human body in white against black background and describe the human activity by a sequence of human body postures, as shown in Figure 3. Depending of the number of available views used for recognition, two methods have been devised, which are presented subsequently.

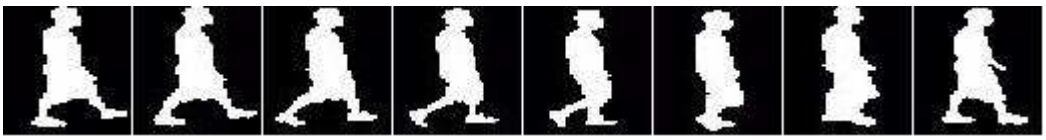

Figure 3: Binary walking human body masks.

Single–view human activity recognition: In the case of single-view movement recognition, binary body masks are extracted from every video frame. Every such mask is centered at the body mass center and rescaled in order to produce binary posture frames of the same size [13]. These are vectorized to produce posture vectors.

In the training phase, posture vectors of the training sequences are clustered to a fixed number of classes using a fuzzy C-means (FCM) algorithm [16]. The resulting cluster centers correspond to the so-called *dyneme* vectors, used in subsequent stages. Each dyneme can be thought of as the average of similarly looking body postures. Since no labelling information is used, the resulting dynemes can represent movement postures appearing in more than one movements. Eight single-view dynemes are presented in Figure 4. After the computation of the dyneme vectors, every posture vector is expressed through its membership vector, which denotes the relationship between a body posture binary mask and the various dynemes.

In order to discriminate movement classes, labelling information available in the training phase is exploited. A linear discriminant analysis (LDA) algorithm is used to project movement vectors in a discriminant subspace. Movement vectors of the various image sequences are projected with LDA and the average of projected movement vectors of all sequences depicting the same movement (e.g. all walking sequences) is computed to represent this movement class.

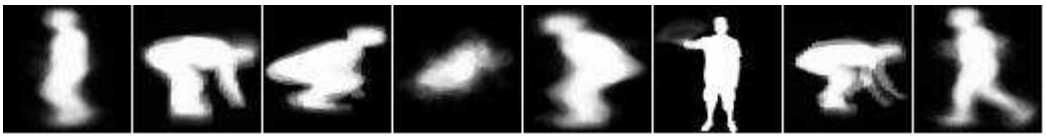

Figure 4: Eight single-view dynemes.

In the recognition phase, binary masks are pre-processed in the same manner with the one used in the training phase. Membership vectors from all the dynemes are calculated and are projected to the discriminant subspace defined by the LDA procedure. Finally, minimum Euclidean or Mahalanobis distance between the projected vector and movement vectors denotes the recognized movement. The obtained results are good, provided that we know the position of the camera vs the human body, which should remain unaltered in the experiments. If this is not the case, then we have to resort to multi-view video sequences.

Multi-view human activity recognition: In the case of multi-view movement recognition, the procedure described above is applied to multi-view videos depicting a person performing a movement from different view-angles [14,15].

In the training phase, single-view binary posture frames are manually ordered, vectorized and concatenated to produce multi-view dynemes with a consistent viewing order, i.e. by placing the frontal view first followed by the other ones in a clockwise manner.

In the recognition phase, the camera correspondence problem, i.e. ordering the cameras according to their location around the person in a consistent manner to the training procedure, is solved with one of the following ways: a) *Correlation based procedure*. Every circular shifted combination of the input posture vectors is compared with every multi-view dyneme found during the training phase. The circular shifted version of

the input posture vectors with the minimum Euclidean distance from a dyneme defines the correct camera ordering. b) *Fourier representation*. By exploiting the circular shift invariance of the Fourier transform representation, posture vectors are represented by their Discrete Fourier Transform. This results to view-independent movement representation.

Ordered single-view posture frames create multi-view posture frames which are then vectorized. The membership vectors are calculated and projected to the discriminant subspace. Finally, the minimum Euclidean or Mahalanobis distance between the projected vector and movement vectors is used to recognize the movement.

FACE/FACIAL EXPRESSION RECOGNITION USING DISCRIMINANT NMF

One of the promising approaches to face recognition and facial expression recognition is to formulate them into a single framework using an appropriate subspace technique, such as the Discriminant Non-negative Matrix Factorization (DNMF) algorithm [17,18], coupled with a modified variant of the Support Vector Machine (SVM) classifier. DNMF is a matrix decomposition algorithm that extends the Non-negative Matrix Factorization (NMF) algorithm. NMF is an unsupervised algorithm that allows only additive combinations of non negative components. DNMF [17,18] was the result of an attempt to introduce discriminant information to the NMF decomposition in a supervised manner. The NMF and DNMF algorithms are briefly presented below.

Let an image scanned row-wise so as to form a vector $\mathbf{x} = [x_1 \ldots x_F]^T$. The basic idea behind NMF is to approximate an image $\mathbf{x}$ (with as small approximation error as possible) by a linear combination of a set of basis images in $\mathbf{Z}$, whose coefficients are the elements of $\mathbf{h}$, such that $\mathbf{x} \approx \mathbf{Zh}$. In order to train the NMF, the matrix $\mathbf{X}$ is constructed, where X_{ij} is the i-th element of the j-th image vector. In other words, the j-th column of $\mathbf{X}$ is the facial image $\mathbf{x}_j$. NMF aims at finding two matrices $\mathbf{Z}$ and $\mathbf{H}$ such that:

$$\mathbf{X} \approx \mathbf{ZH}. \qquad (1)$$

Obviously, the application of NMF requires the evaluation of the basis images in $\mathbf{Z}$. This is done at a training phase that requires a set of training images $\mathbf{x}_1 \ldots \mathbf{x}_n$, where n is the training set size. After the NMF decomposition, the facial image $\mathbf{x}_j$ can be written as $\mathbf{x}_j \approx \mathbf{Zh}_j$, where $\mathbf{h}_j$ is the j-th column of $\mathbf{H}$. Thus, the columns of the matrix $\mathbf{Z}$ can be considered as the basis images and the vector $\mathbf{h}_j$ as the weight vector that corresponds to image $\mathbf{x}_j$. The vector $\mathbf{h}_j$ can be also considered as the projection of $\mathbf{x}_j$ to a lower dimensional space. The cost for the decomposition (1) can be defined as the sum of all KL divergences for all images in the database:

$$\mathrm{D}\left(\mathbf{X} \| \mathbf{ZH}\right) = \sum_j \mathrm{KL}\left(\mathbf{x}_j \| \mathbf{Zh}_j\right) \qquad (2)$$

To formulate the DNMF method, discriminant constraints have been incorporated in the NMF decomposition, inspired by the minimization of Fisher's criterion [15]. The matrix $\mathbf{S}_w$ defines the scatter of the sample vector coefficients around their class mean. The dispersion of samples that belong to the same class around their corresponding mean should be as small as possible. A convenient measure for the dispersion of the samples is the trace of $\mathbf{S}_w$. The matrix $\mathbf{S}_b$ denotes the between-class scatter matrix and defines the scatter of the mean vectors of all classes around the global mean. Each class must be as far as possible from the other classes. Therefore, the trace of $\mathbf{S}_b$ should be as large as possible. Thus, the DNMF cost function is given by:

$$\mathrm{D}_d\left(\mathbf{X} \| \mathbf{ZH}\right) = \mathrm{D}\left(\mathbf{X} \| \mathbf{ZH}\right) + \gamma \mathrm{tr}\left[\mathbf{S}_w\right] - \delta \mathrm{tr}\left[\mathbf{S}_b\right] \qquad (3)$$

where γ and δ are non-negative constants. The update rules that guarantee a non-increasing behavior of (3) for the weights and the base images, under the non negative constraints, can be found in [17].

In order to form the training and test sets for the proposed framework, face detection and tracking were applied on the video frames containing faces. The resulting Regions Of Interest (ROIs) were anisotropically scaled, so as to have fixed size of 30 × 40 pixels and were converted to grayscale facial images. Each such fixed size facial image was scanned row-wise, so as to form a feature vector $\mathbf{x} = [f_1 \ldots f_{1200}]^T$ (f_i being the luminance of the i-th pixel), which was used to compose the training and test sets for face and facial expression recognition, respectively. Different training image sets have been used for each of these two tasks. For each task, the resulting training set is fed to DNMF algorithm. During training, the basis images **Z** are calculated by the application of DNMF on the training face images. During testing, a facial image under examination is firstly projected to the derived lower dimensional feature space $\mathbf{g} = \mathbf{Z}^T \mathbf{x}$ and is later inserted into the SVM system that performs classification into one of the predefined, during training, face or facial expression classes.

Extensive experimentation has been performed for face recognition on the XM2VTS database [17] and for facial expression recognition on the Cohn-Kanade database [18]. The results demonstrate that the proposed framework achieves high recognition accuracy rates for both problems.

CONCLUSIONS AND FUTURE WORK

We have presented a framework for anthropocentric digital video analysis that is particularly suited, among others, to film and games postproduction, indexing and retrieval as well as to the semantic labelling of audiovisual material. The presented framework is far from complete, since it does not cover all relevant topics, e.g. facial feature (eye/mouth) recognition, visible speech detection, facial video summarization, uncalibrated 3D face/head/body reconstruction, to name a few. The interested reader can find some related literature in [1]. The entire field has many different and very interesting research topics that may be trivial for humans but very difficult for machines, e.g. the study of human interactions and behaviour.

ACKNOWLEDGEMENT

The research leading to these results has received funding from the European Community's Seventh Framework Programme (FP7/2007-2013) under Grant Agreement 211471 (i3DPost).

REFERENCES

[1] N.Vretos , V.Solachidis and I.Pitas , "Anthropocentric semantic information extraction from movies" in Computational Intelligence in Multimedia Processing: Recent Advances, edited by: Aboul-Ella Hassanien Janusz Kacprzyk and Ajith Abr, Springer Publisher Co., 2007.

[2] Nikolaos Tsapanos, Anastasios Tefas, Ioannis Pitas: Online shape learning using binary search trees. Image Vision Comput. 28(7): 1146-1154 (2010).

[3] Dariu Gavrila: A Bayesian, Exemplar-Based Approach to Hierarchical Shape Matching. IEEE Trans. Pattern Anal. Mach. Intell. (PAMI) 29(8):1408-1421 (2007).

[4] Navneet Dalal, Bill Triggs: Histograms of Oriented Gradients for Human Detection. CVPR 2005:886-893.

[5] Zhe Lin, Larry S. Davis: Shape-Based Human Detection and Segmentation via Hierarchical Part-Template Matching. IEEE Trans. Pattern Anal. Mach. Intell. (PAMI) 32(4):604-618 (2010)

[6] Luca Marchesotti, Gianni Vernazza, and Carlo S. Regazzoni, "A multicamera fusion framework for multiple occluding objects tracking in intelligent monitoring and sport viewing applications," in IEEE ICIP, 2004, pp. 1033–1036.

[7] Abhinav Gupta, Anurag Mittal, and Larry S. Davis, "Constraint integration for multiview pose estimation of humans with self-occlusions," in Proceedings of the Third International Symposium on 3D Data Processing, Visualization, and Transmission (3DPVT'06), Washington, DC, USA, 2006, pp. 900–907.

[8] Ming Xu, James Orwell, and Graeme Jones, "Tracking football players with multiple cameras," in IEEE ICIP, 2004, pp. 2909–2912.

[9] James Black, Tim Ellis, and Paul Rosin, "Multi view image surveillance and tracking," in MOTION '02: Proceedings of the Workshop on Motion and Video Computing, Washington, DC, USA, 2002, pp. 169–174.

[10] G. Sfiris, N. Nikolaidis, I. Pitas "Multi-view Object and Human Body Part Detection Utilizing 3D Scene Information", Proceedings, IEEE ICIP 2010, Hong Kong, 26-29/9/2010.

[11] Paul Viola and Michael J. Jones, "Robust real-time face detection," International Journal of Computer Vision, vol. 57, no. 2, pp. 137–154, 2004.

[12] G.Stamou , M.Krinidis , N.Nikolaidis and I.Pitas , "A monocular system for automatic face detection and tracking" in Proc. of Visual Communications and Image Processing (VCIP 2005), Beijing, China, 12-15 July, 2005.

[13] N. Gkalelis, A. Tefas, and I. Pitas, "Combining fuzzy vector quantization with linear discriminant analysis for continuous human movement recognition," IEEE Trans. Circuits Syst. Video Technol., vol. 18, no 11, pp. 1511-1521, Nov. 2008.

[14] N.Gkalelis, N.Nikolaidis and I.Pitas , "View independent human movement recognition from multi-view video exploiting a circular invariant posture representation" in Proc. IEEE International Conference on Multimedia and Expo (ICME 2009), New York, USA, June - July, 2009.

[15] A.Iosifidis, N.Nikolaidis and I.Pitas , "Movement recognition exploiting multi-view information" Proceedings of the IEEE International Workshop on Multimedia Signal Processing (MMSP 2010), St. Malo, France, 4-6/10/2010.

[16] J. Bezdek, Pattern Recognition with Fuzzy Objective Function Algorithms. New York: Plenum, 1981.

[17] S. Zafeiriou, A. Tefas, I. Buciu and I. Pitas, "Exploiting Discriminant Information in Non-negative Matrix Factorization with application to Frontal Face Verification" IEEE Transactions on Neural Networks, vol. 17, no. 3, pp. 683-695, May, 2006.

[18] I. Buciu and I. Pitas, "NMF, LNMF and DNMF modeling of neural receptive fields involved in human facial expression perception" Journal of Visual Communication and Image Representation, vol. 17, no. 5, October, 2006.

ABOUT THE AUTHORS

Professor Ioannis Pitas, PhD, Department of Informatics, Aristotle University of Thessaloniki, Greece, also with the Informatics and Telematics Institute, CERTH, Greece, Phone: +30-2310-996304, E-mail: pitas@aiia.csd.auth.gr.

Assistant Professor Nikolaos Nikolaidis, PhD, Department of Informatics, Aristotle University of Thessaloniki, Greece, also with the Informatics and Telematics Institute, CERTH, Greece, Phone: +30-2310-998566, E-mail: nikolaid@aiia.csd.auth.gr.