Using Mutual Information to Indicate Facial Poses in Video Sequences

Georgios Goudelis
Aristotle University of Thessaloniki
Thessaloniki, Greece
goudelis@aiia.csd.auth.gr

Anastasios Tefas
Aristotle University of Thessaloniki
Thessaloniki, Greece
tefas@aiia.csd.auth.gr

Ioannis Pitas
Aristotle University of Thessaloniki
Thessaloniki, Greece
pitas@aiia.csd.auth.gr

ABSTRACT

Estimation of the facial pose in video sequences is one of the major issues in many vision systems such as face based biometrics, scene understanding for human and others. The proposed method uses a novel pose estimation algorithm based on mutual information to extract any required facial pose from video sequences. The method extracts the poses automatically and classifies them according to view angle. Experimental results on the XM2VTS video database indicated a pose classification rate of 99.2% while it was shown that it outperforms a PCA reconstruction method which was used as a benchmark.

1. INTRODUCTION

Facial pose is one of the major issues concerning surveillance systems based on human behavior and intentions, as well as for face based biometric applications. Facial pose estimation from video sequences is a task of great importance for vision systems performing scene understanding for human-computer interfaces or security surveillance [15]. A number of works can be found in bibliography that attempt to estimate facial pose or to use this information for a number of different applications.

In [15] an analysis of face similarity distributions under varying head pose for different type of image transformation with the aim of understanding pose in similarity space is presented. In this work, the use of Gabor filters and PCA as transformation of prototypes images in order to emphasize pose differences is examined. In [4] a deformable graph is used to determine face position and pose from learned models. However, the method is highly time consuming and is not appropriate for real-time applications. In [11], the work on eigenfaces is extended to modular eigenspaces in order to estimate the pose of a face while in [7] the combination of support vector regression and modular Support Vector Machines (SVMs) was used for pose estimation and face detection respectively.

Other approaches use video in order to solve the pose estimation problem or to take advantage from pose extraction for application such as face-based biometrics. More specifically, in [5], each registered person is represented by a low-dimensional appearance manifold in the ambient image space. This manifold is approximated by piecewise linear subspaces and the dynamics among them are embodied in a transition matrix learned from an image sequence. In [8], a method for real-time multi-view face detection and facial pose estimation is described. The method employs a convolutional network to map face images to points on a manifold parameterized by pose and non-face images to points far from manifold. The network is trained by optimizing a loss function of three variables: image, pose and face/non-face label. Finally, in [6] an Independent Component Analysis (ICA) based approach is presented for learning view-specific subspace representations of the object from multiview face examples. Two variants of ICA, namely Independent Subspace Analysis (ISA) and topographic Independent Component Analysis (TICA), take into account higher order statistics needed for object view characterization. ICA, TICA and ISA are proven to learn view-specific basis components from the mixture data.

In this paper, we propose a novel method for automatic pose extraction in head-and-shoulder videos. The method is able to find any pose required. It is based on mutual information and evaluates the information content of each facial image (contained in a video frame) of facial poses in comparison to a given ground truth image. The experiments produced a pose classification rate of 99.2% on all examined pose cases.

2. FACIAL POSE ESTIMATION

Mutual information has been previously used in computer vision, for example in image registration [17] or in audio-visual speech acquisition [12]. The mutual information of two random variables measures the mutual dependence of the two variables [13], [14], [9]. In our case, mutual information measures the dependency of the information contained in two video frames. The closer the mutual information between two frames is to zero, the less information one frame contains about the other and vise versa.

To the best of the author's knowledge, this is the first work that uses the manifold that is defined by the MI distance in order to measure the similarity between facial poses. The method performed very well in our experimental procedure and appears to be robust against small changes in scale and

illumination.

Let X be a discrete random variable with a set of possible outcomes $\mathcal{A}_X = \{a_1, a_2, \ldots, a_N\}$ having probabilities $\{p_1, p_2, \ldots, p_N\}$, with $p_X(x = a_i) = p_i, p_i \geq 0$ and $\sum_{x \in \mathcal{A}_X} p_X(x) = 1$. Entropy measures the information content or "uncertainty" of X and it is given by [1]:

$$H(X) = -\sum_{x \epsilon \mathcal{A}_X} p_X(x) \log p_X(x). \quad (1)$$

The *joint entropy* is a statistic measure that summarizes the degree of dependency of random variable X on random variable Y. The *joint entropy* of X, Y is expressed as:

$$H(X,Y) = -\sum_{x,y \in \mathcal{A}_X, \mathcal{A}_Y} p_{XY}(x,y) \log p_{XY}(x,y) \quad (2)$$

where $p_{XY}(x,y)$ is the joint probability density function. For two random variables X and Y, the *conditional entropy* of Y given X is denoted as $H(Y \mid X)$ and is defined as:

$$\begin{aligned} H(Y \mid X) &= \textstyle\sum_{x \in \mathcal{A}_X} p_X(x) H(Y \mid X = x) = \\ &= -\textstyle\sum_{x,y \in \mathcal{A}_X, \mathcal{A}_Y} p_{XY}(x,y) \log p_{XY}(x \mid y) \end{aligned} \quad (3)$$

where $p_{XY}(x \mid y)$ denotes the conditional probability. The conditional entropy $H(Y \mid X)$ is the uncertainty in Y, given knowledge of X. It specifies the amount of information that is gained by measuring a variable when already knowing another one. It is very useful if we want to know whether there is a functional relationship between two data sets (e.g. two facial image regions). The mutual information between the random variables X and Y is given by:

$$I(X,Y) = -\sum_{x,y \in \mathcal{A}_X, \mathcal{A}_Y} p_{XY}(x,y) \log \frac{p_{XY}(x,y)}{p_X(x) p_Y(y)} \quad (4)$$

and measures the amount of information conveyed by X about Y. The relation between the mutual information and the joint entropy of random variables X and Y is given by:

$$I(X,Y) = H(X) + H(Y) - H(X,Y), \quad (5)$$

where $H(X)$ and $H(Y)$ are the marginal entropies of X and Y. The mutual information is a measure of the additional information known about X when Y is given:

$$I(X,Y) = H(X) - H(X \mid Y), \quad (6)$$

where $H(X)$ is the marginal entropy, $H(X \mid Y)$ is conditional entropy, and $H(X,Y)$ is the joint entropy of X and Y. According to (5), the mutual information provides us with a measure of correspondence between X and Y. We can also see from (6) that the mutual information is reduced, if X carries no information about Y.

2.1 Calculation of Mutual Information between video frames

In order to calculate the mutual information between two video frames we have to use a reference frame and a test frame denoted by X and Y accordingly, having L pixels each. We can consider pixel values as outcomes $\mathcal{A}_X = \{a_1, a_2, \ldots, a_L\}$ of a random variable. The probabilities p_X, p_Y in (4), are estimated by the histograms of the images U and V, while the joint probability p_{XY} is estimated by the joint histogram of the images [16]. The density probabilities of both images are estimated using the Parzen Window technique [10], [2]. This is a classical technique used in neural networks for estimating a probability density function (pdf) from a sample.

While entropy for an image remains fixed, joint entropy and mutual information of two images vary as the 1-1 correspondence between the pixels from each image changes with every geometrical alignment. When mutual information is maximized, the geometric relationship, under which one image explains the other most effectively, is achieved. In other words, the maximization of mutual information provides image registration.

2.2 Pose Classification

Most videos are obtained using an uncalibrated camera. This means that we do not know the three angles that define a persons pose with respect to the camera reference system. Therefore, we consider pose estimation as a classification problem in which we assign a facial image to a particular pose class (e.g. frontal, left/right profile) by examining its similarity with other images of the video. In order to describe how the pose is assigned to a class, let us describe the enrolment procedure for a candidate reference person r. The system contains an image database $\mathcal{C} = \bigcup_r C_r$, where $\mathcal{C}_r$ is the set of images assigned to the reference person r. Suppose that $\mathcal{C}_{jr}$ is subset of the $\mathcal{C}_r$ that contains the images of j-th pose of person r so that $\mathcal{C}_r = \bigcup_j \mathcal{C}_{jr}$. Let the video V_r be partitioned to the set of N frames $F_{1r}, ..., F_{Nr}$. Each video frame F_{kr} (where $k = 1, \ldots, N$), is examined and compared with the images of the set $\mathcal{C}_r$ in order to assign it to a pose set $\mathcal{C}_{jr}$. The mean mutual information:

$$I_m(F_{kr}, \mathcal{C}_{jr}) = \frac{1}{\mathcal{N}(\mathcal{C}_{jr})} \sum_{\mathbf{c}_i \in \mathcal{C}_{jr}} I(F_{kr}, \mathbf{c}_i) \quad (7)$$

is calculated using equation (6) for every $F_{kr} \in V_r$. $\mathbf{c}_i$ is an image of $\mathcal{C}_{jr}$ and $\mathcal{N}(C_{jr})$ denotes the cardinality of the set C_{jr}. The frame F_{kr} is assigned to the pose class corresponding to the maximum $I_m(F_{kr}, \mathcal{C}_{jr})$:

$$pose = \arg\max_j I_m(F_{kr}, \mathcal{C}_{jr}). \quad (8)$$

2.3 Description of experiments and results

In order to make the procedure more clear to the reader and present a possible use of the proposed method, we set forth a hypothetical scenario that describes a verification system based on the proposed face detection algorithm, as a real world application.

A surveillance camera is installed in a building where special security issues are required and comprise part of the security system. In the specific system a number of persons have been enroled. During the enrolment procedure, the person is asked to turn its head in all possible directions in front of a recording camera in order to store in a database every facial pose of the person. For some poses that seem to be rich in information content like frontal, right and left profiles, a pose specific training procedure is followed. This way all the crucial facial poses of a client are learned by the system. Supposing that a person is requiring access to the building (i.e. an identity claim occurs). Unlike the enrolment procedure which is supervised the testing procedure

is fully automatic. The surveillance camera is recording the scene while a face detector locates the facial area in every frame. The pose classes are used according to the recognition/verification algorithm that is used and a decision is taken for the acceptance or the rejection of the claim.

2.4 Evaluation of the pose detection algorithm

For testing the capability of the algorithm having many persons and different video sessions of them, we used the XM2VTS video database. This database, contains four recordings of 295 subjects taken over a period of four months. Each recording is comprised of a speaking head shot and a rotating head shot. Sets of data taken from this database are available including high quality color images, 32 KHz 16-bit sound files, video sequences and a 3D model. In the first shot of each session the persons read a given text, while in the second shot each person moves its head in all possible directions allowing multiple views (poses) of the face.

In our framework, video is processed frame by frame. Only the subsampled luminance is used in order to reduce processing time. Afterwards, the uniform background is removed using a grassfire algorithm [3]. To achieve a better and more accurate verification rate, the algorithm resizes each video, according to a factor produced by a given standard distance between right and left eye, when the subject is in frontal position, while it keeps the frame size stable. This way, the scaling problem occurred in different sessions of the same person, is resolved (at least partially), while the head is aligned for the frame representing the frontal face and, consequently, for the most part of the remaining of the movement. A sample sequence of already processed frames is illustrated in Figure 1.

Figure 1: Sample of XM2VTS video sequence. The person starts from the frontal pose, turn its head from right to left and returns to frontal pose.

For each of the persons examined, a ground truth image representing the required pose is used. This image is always taken from a different session from the one examined. The ground truth constitutes F_{kr} in equation (7), while $\mathbf{c}_i$ is every following frame of the examined video input. This way, each frame is compared to the ground truth and their mutual information is stored in a vector $\mathbf{q}$.

The plot of this vector entries $\mathbf{q}_i$, (for $i = 1, \ldots, N$) for a smooth movement is given in Figure 2. The mutual information in this case has been calculated using the frontal face of the person taken from a different session than the one under examination.

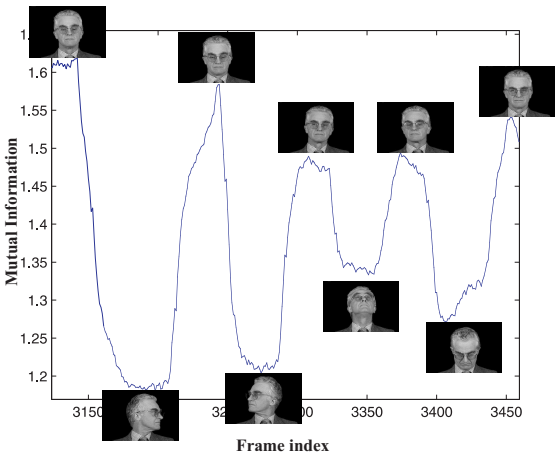

Figure 2: Mutual information plot for a smooth head movement vs video frame time index. The characteristic facial image poses are superimposed.

The plot clearly shows how mutual information changes, as the head passes from pose to pose. The mutual information is maximized when the head pose is close to the frontal position. This clearly shows how the mutual information "detects" the similarity between the first frame which represents the frontal face and every near to frontal pose that appears within the video sequence.

2.4.1 *Experiments on aligned scaled frames*

Likewise, we repeat the same procedure using specific poses extracted from a different video session than the test one. The method was tested on 120 different persons using each time (as ground truth), images taken from each of the four different sessions. Thus, 4320 (4 sessions $\times$ 3 poses $\times$ 3 tests to every session $\times$ 120 persons) testings where carried out. The result obtained was of high accuracy showing that the algorithm was able to find correctly 99.2% of the required poses. Some successful matchings are presented in Figure 3.

Figure 3: Successful pose estimations. a) ground truth images, b) results produced

2.4.2 *Comparison of pose detection algorithm with PCA reconstruction method*

To make a comparison of the pose detection algorithm with a method that could be used as a benchmark, we used a

Table 1: Percentages (%) of correct facial pose classification

Method	Right	Mid-Right	Frontal	Mid-Left	Left	Up	Down
PCA	75.2	63.3	85.5	65.7	76.0	79.7	81.6
Proposed	85.1	65.3	90.2	71.4	87.1	78.3	80.1

PCA reconstruction method for pose classification. For each pose class, a PCA model was constructed. The model with the smaller reconstruction error was the one finally classified (i.e., smallest L_2 norm distance). This scenario is like using eigenfaces method for every pose and when an unknown pose arrives, the test image is projected to all different pose-subspaces. The one obtaining the minimum L_2 distance is the winner. The experiment was performed for both the above described databases. PCA was trained with ground truth facial pose images extracted by the video sequences. The experiment has run on different sessions for different bounding box dimensions and no alignment or rescaling has taken place. Due to space limitations we report only the best results for both methods produced by the use of bounding boxes with sizes 77×69. The results of the comparison of the two methods, are given in Table 1.

It can be easily seen that the performance of the proposed method is significantly better for almost every pose examined. On average, the proposed method outperforms PCA reconstruction method by 7.89%.

3. CONCLUSION

In this paper, a novel way for automatic facial pose extraction on mutual information is proposed. The information in video frames is compared with the information contained in a ground truth image representing the required pose. Experimental results on the XM2VTS video database, show that the algorithm is able to perform very well for a number of different requested poses when tracker information is used. The method proved to outperform a PCA reconstruction method which was used as a benchmark. It is worth to be noted, that the pose detection algorithm proved to be robust in small variations of scale and illumination.

4. REFERENCES

[1] T. M. Cover and J. A. Thomas. *Elements of information theory*. Wiley-Interscience, New York, NY, USA, 1991.

[2] R. Duda and P. Hart. *Pattern Classification and Scene Analysis*. John Wiley and Sons, 1973.

[3] C. Kotropoulos, A. Tefas, and I. Pitas. Morphological elastic graph matching applied to frontal face authentication under well-controlled and real conditions. *Pattern Recognition*, 33(12):31–43, Oct. 2000.

[4] N. Kruger, M. Potzsch, and C. von der Malsburg. Determination of face position and pose with a learned representation based on labeled graphs. In *Image and Vision Computing*, pages 665–673, 1997.

[5] K.-C. Lee, J. Ho, M.-H. Yang, and D. J. Kriegman. Video-based face recognition using probabilistic appearance manifolds. In *CVPR*, pages 313–320, 2003.

[6] S. Z. Li, X. Lu, X. Hou, X. Peng, and Q. Cheng. Learning multiview face subspaces and facial pose estimation using independent component analysis. *IEEE Transactions on Image Processing*, 14(6):705–712, 2005.

[7] Y. Li, S. Gong, and H. Liddell. Support vector regression and classification based multi-view face detection and recognition. In *FG: Proceedings of the Fourth IEEE International Conference on Automatic Face and Gesture Recognition 2000*, page 300. IEEE Computer Society, 2000.

[8] R. Osadchy, M. Miller, and Y. LeCun. Synergistic face detection and pose estimation with energy-based model. In *Advances in Neural Information Processing Systems (NIPS 2004)*. MIT Press, 2005.

[9] A. Papoulis. *Probability, Random Variables, and Stochastic Processes*. pub-mcgraw-hill, 3rd edition, 1991.

[10] E. Parzen. On the estimation of a probability density function and mode. *Annals of Mathematical Statistics*, 33:1065–1076, 1962.

[11] A. Pentland, B. Moghaddam, and T. Starner. View-based and modular eigenspaces for face. In *In IEEE Conference on Computer Vision and Pattern Recognition (CVPR'94)*, Seattle, WA, June 1994.

[12] D. Roy and A. Pentland. Learning words from natural audio-visual input. *In International Conference on Spoken Language Processing*, 4:1279–1283, 1999, Sydney, Australia.

[13] C. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27:379–423 and 623–656, July and October 1948.

[14] C. Shannon and W. Weaver. *Mathematical Theory of Communication*. University of Illinois Press, June 2002.

[15] J. Sherrah, S. Gong, and E. Ong. Face distribution in similarity space under varying head pose. *Image and Vision Computing*, 19(11), 2001.

[16] P. Thevenaz and M. Unser. An efficient mutual information optimizer for multiresolution image registration. In *Proceedings of the 1998 IEEE International Conference on Image Processing (ICIP'98)*, volume I, pages 833–837, Chicago IL, USA, October 4-7, 1998.

[17] P. A. Viola. Alignment by maximization of mutual information. Technical Report AITR-1548, 1995.