Inexact Graph Matching for Facial Feature Segmentation and Recognition in Video Sequences: Results on Face Tracking

Ana Beatriz V. Graciano[1], Roberto M. Cesar Jr.[1], and Isabelle Bloch[2]

[1] Department of Computer Science, IME, University of São Paulo. São Paulo, Brazil.
{cesar,abvg}@ime.usp.br
[2] Signal and Image Processing Department, CNRS UMR 5141 LTCI, École Nationale
Supérieure des Télécommunications. Paris, France.
Isabelle.Bloch@enst.fr

Abstract. This paper presents a method for the segmentation and recognition of facial features and face tracking in digital video sequences based on inexact graph matching. It extends a previous approach proposed for static images to video sequences by incorporating the temporal aspect that is inherent to such sequences. Facial features are represented by *attributed relational graphs*, in which vertices correspond to different feature regions and edges to relations between them. A reference model is used and the search for an optimal homomorphism between its corresponding graph and that of the current frame leads to the recognition.

1 Introduction

This paper deals with segmentation and recognition of facial features in digital video sequences through the use of an inexact graph matching technique. The proposed technique constitutes a first approach to the generalization of the methodology developed in [3,4] for facial feature segmentation and recognition in static images.

This extension is motivated by the fact that the subject of face analysis and recognition arises in various computer vision applications involving human activity recognition, such as affective computing, surveillance, teleconferencing and multimedia databases. Since many of these involve video sequence processing, it is also interesting to incorporate the notion of *motion-based recognition* [8] in the methodology.

The main idea is to model the target facial features in a given face through an *attributed relational graph* (ARG), a structure in which vertices represent the facial features and their attributes, while edges represent spatial relations among them. The model image is manually segmented into the facial features of interest and relations are computed to derive a model ARG. In each input image where recognition has to be performed, i.e. each frame, its gradient is extracted and a watershed algorithm is applied to it. Then, an input ARG is obtained from the resulting oversegmented image as for the model. The recognition step relies on an

A. Sanfeliu and J. Ruiz-Shulcloper (Eds.): CIARP 2003, LNCS 2905, pp. 71–78, 2003.

inexact graph matching procedure that finds a suitable homomorphism between the graph obtained from the model and the one obtained from the image.

The technique of inexact graph matching has been extensively studied in several different domains [2,9,10], such as pattern recognition, computer vision, cybernetics, among others. This approach is justified here due to the difficulty in finding an isomorphism between the model image graph and the input image one: since the latter represents an oversegmented image, it is not possible to expect a bijective correspondence between both structures.

It is worth noting that the term facial feature recognition used hereby means that each facial feature of interest will be located and classified as such. Therefore, it is not related to the recognition performed as a means of matching a face against a known database of faces for instance (no face recognition is performed). Based on [3] where the static methodology is introduced and on [4] where the optimization of the graph matching process is addressed using several methods, the main contribution of the present work is to develop a methodology that can be applied to video sequences, i.e. incorporating the temporal dimension.

This paper is organized as follows. Section 2 explains how a face is modeled as an attributed relational graph. Section 3 explains the inexact graph matching step of the methodology. Section 4 shows how the tracking process is performed throughout the video frames. Section 5 presents some obtained results and conclusions.

2 Face Representation

Attributed Relational Graphs. In this work, a directed graph will be denoted by $\tilde{G} = (N, E)$, where N represents the set of vertices of $\tilde{G}$ and $E \subseteq N \times N$ the set of edges. Two vertices a, b of N are said to be adjacent if $(a, b) \in E$. When each vertex of $\tilde{G}$ is adjacent to all others, then $\tilde{G}$ is said to be complete. Furthermore, $|N|$ denotes the number of vertices in G, while $|E|$ denotes its number of edges.

An *attributed relational graph* (also referred to as ARG) is a graph in which attribute vectors are assigned to vertices and to edges. Formally, we define an ARG as $G = (N, E, \mu, \nu)$, where N represents the set of vertices of G and $E \subseteq N \times N$ the set of edges. Furthermore, $\mu : N \rightarrow L_N$ assigns an attribute vector to each vertex of G, while $\nu : E \rightarrow L_E$ assigns an attribute vector to each edge in G.

The structure of a face can be thought of as being a collection of features (e.g: lips, eyebrows, nostrils, chin) which are somehow related in terms of their relative positions on the face. In the proposed model, facial feature regions are represented by vertices in a graph, while relations between them are represented by edges. The attribute vectors μ and ν may also be called object and relational attributes, respectively. The former refers to connected regions in the image and the latter to the spatial arrangement of the regions.

Attributes. The object and relational attributes convey the knowledge about faces to the ARG structure. The attributes which have been considered in this work are the same as in [3]. Let us consider an ARG $G = (N, E, \mu, \nu)$ and any two vertices a, b in N.

The object attribute $\mu(a)$ is defined as:

$$\mu(a) = (g(a), w(a), l(a)). \tag{1}$$

The term $g(a)$ corresponds to the average gray-level of the image region associated to vertex a, whereas $w(a)$ is a coefficient obtained from the application of a Morlet wavelet. Both $g(a)$ and $w(a)$ are normalized between 0 and 1 with respect to the maximum possible grey-level. Finally, $l(a)$ is a region label.

The relational attribute $\nu(a, b)$, for a, b in E, is defined as:

$$\nu(a, b) = (\overrightarrow{v}, sym(a, b)). \tag{2}$$

The first attribute is the vector $\overrightarrow{v} = (p_b - p_a)/2d_{max}$, where d_{max} is the maximum distance between any two points of the input graph, while p_a and p_b denote the centroids of the image regions to which vertices a and b correspond. The term $sym(a, b)$ denotes a reflectional symmetry calculated as described in [1].

The Face Model. A face model image is used as a reference to recognize facial features of interest. This image can be for instance the first frame of a given video sequence. It is manually segmented into facial feature regions of interest and the landmark of each region is calculated. Then, the corresponding ARG is derived.

The model graph should contain vertices associated to each target facial feature region (e.g. lips, iris, eyebrows, skin). However, if a single feature presents considerable variability within its domain, it might need to be subdivided into smaller sub-regions, so that the averages considered when calculating both vertex and edge attributes can be more representative.

3 The Facial Feature Recognition Process

Graph Homomorphism. Consider two ARGs $G_1 = (N_1, E_1, \mu_1, \nu_1)$ derived from the image and $G_2 = (N_2, E_2, \mu_2, \nu_2)$ derived from the model. They will be called *input* and *model* graphs respectively. Also, subscripts will be used to refer to vertices and edges in each graph, e.g. $a_1 \in N_1$ is a vertex in G_1, $(a_2, b_2) \in E_2$ is an edge in G_2. It is also important to notice that, since G_1 results from an oversegmented image, $|N_1|$ is much greater than $|N_2|$ in general.

An *association graph* $\tilde{G}_A$ between G_1 and G_2 is defined as the complete graph $\tilde{G}_A = (N_A, E_A)$, where $N_A = N_1 \times N_2$ and $E_A = E_1 \times E_2$.

A graph homomorphism h between G_1 and G_2 is a mapping $h\colon N_1 \to N_2$ such that $\forall a_1 \in N_1, \forall b_1 \in N_1, (a_1, b_1) \in E_1 \Rightarrow (h(a_1), h(b_1)) \in E_2$. This definition assumes that all vertices in G_1 should be mapped to G_2.

Finding a homomorphism between G_1 and G_2 is essential to the face feature recognition process. Since $|N_1|$ is greater than $|N_2|$, a suitable homomorphism between the input and model graphs should map distinct vertices of G_1 into a single vertex of G_2, which corresponds to merging coherent sub-regions in the input oversegmented image.

As proposed in [3], a solution for finding a homomorphism between G_1 and G_2 may be defined as a complete sub-graph $\tilde{G}_S = (N_S, E_S)$ from the association graph $\tilde{G}_A$, in which $N_S = \{(a_1, a_2), a_1 \in N_1, a_2 \in N_2\}$ such that $\forall a_1 \in N_1, \exists a_2 \in N_2, (a_1, a_2) \in E_S$, and $\forall (a_1, a_2) \in E_S, \forall (a_1', a_2') \in E_S, a_1 = a_1' \Rightarrow a_2 = a_2'$, assuring that each vertex from the input graph corresponds to exactly one vertex of the model graph and $|N_S| = |N_1|$. It should be clear that such a solution only considers the structures of G_1 and G_2, and that it gives rise to many possible homomorphisms between both graphs.

Objective Function. In order to evaluate the quality and suitability of a given homomorphism between the input and model graphs, an objective function must be defined. It should consider not only the structure of the graphs, but also the attributes of the facial features and their relations. In this paper, the assessment of a certain homomorphism is accomplished through the minimization of the following function:

$$f(\tilde{G}_S) = \frac{\alpha}{|N_S|} \sum_{(a_1,a_2) \in N_S} c_N(a_1, a_2) + \frac{(1-\alpha)}{|E_S|} \sum_{e \in E_S} c_E(e) \tag{3}$$

where c_N and c_E are dissimilarity measures given as follows:

$$c_N(a_1, a_2) = \left\{ \begin{array}{l} \gamma_N |g_1(a_1) - g_2(a_2)| + (1 - \gamma_N)|w_1(a_1) - w_2(a_2)|, \\ \\ \text{if } l(a_1) = l(a_2) \\ \\ \infty, \text{otherwise} \end{array} \right\} \tag{4}$$

$$c_E(e) = \gamma_E \phi_v + (1 - \gamma_E)\phi_{sym} \tag{5}$$

and ϕ_v and ϕ_{sym} are defined as

$$\begin{aligned} \phi_v &= \gamma_v \big| \|\overrightarrow{v}_1\| - \|\overrightarrow{v}_2\| \big| + (1 - \gamma_v)\tfrac{|\cos\theta - 1|}{2} \\ \phi_{sym} &= |sym(a_1, b_1) - sym(a_2, b_2)| \ . \end{aligned} \tag{6}$$

In this case, $\cos\theta = \frac{\overrightarrow{v}_1 \overrightarrow{v}_2}{\|\overrightarrow{v}_1\| \|\overrightarrow{v}_2\|}$ and the values γ_N, γ_E and γ_v are weighting parameters.

Performing Inexact Graph Matching. Many approaches have been explored for optimizing inexact graph matching for pattern recognition purposes, such as those mentioned in [3], [7] and [10].

In this work, the matching is achieved through the use of a tree search algorithm. Other possible alternatives include genetic and estimation of distribution algorithms [3].

In general terms, the tree-search optimization algorithm builds a search-tree where each vertex represents a pair of vertices (k, l), $k \in N_1$ and $l \in N_2$. The root vertex is labelled $(0, 0)$ and it is expanded in $|N_2|$ sons labelled $(1, l)$, $l = 1...|N_2|$. At each step k of the algorithm, the son which minimizes the objective function, say (k, l_{min}) is chosen and therefore expanded in $|N_2|$ sons $(k+1, i)$, $i = 1...|N_2|$. The process is repeated until a vertex $(|N_1|, l)$ is reached, which guarantees that all vertices of G_1 have been assigned to a vertex of G_2, thus establishing a homomorphism between the input and model graphs.

4 The Tracking Process

In this section, we aim at generalizing the previous approach to video sequences. Since digital video is composed of a sequence of images which change over time, it is needed to incorporate in the methodology this temporal aspect and transitions between images, reflecting facial feature changes throughout the video (e.g. a progressive smile or a blink). In this section, we present our first approach towards reflecting such changes in the facial features.

General Scheme. The overall sequence of steps performed in order to segment and recognize the facial features of interest in a generic frame of the input video sequence is illustrated in Figure 1.

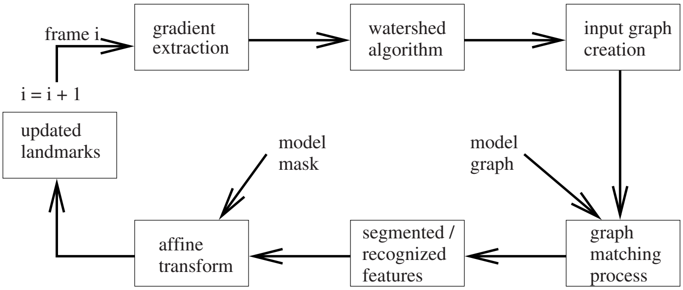

Fig. 1. Overview of the tracking process.

Initially, approximative landmarks of the target facial features are located in the first frame for future constraint on the region in which the oversegmentation will be performed. They are obtained through the use of the Gabor Wavelet Network (GWN) [6]. Then the previous algorithm for static images is applied in the regions of the face around the landmarks.

One of the main contributions in the methodology is related to updating the landmarks which will be used in the subsequent frame in the video sequence, thus avoiding the need for a global face tracking procedure (i.e. GWN in our specific case) in addition to the graph matching. If the same landmarks were applied to all frames, the matching and recognition results would possibly not be satisfactory, since the features in each frame usually have considerable differences in terms of their positions.

Landmark Updating. The GWN technique could be applied to each frame of the sequence in order to update the landmarks. However, it would be interesting to make use of the information obtained directly from the graph methodology and the model image. To accomplish this, an *affine transformation* is applied in order to map the model image to the frame under consideration based on the recognized facial features, allowing the landmark updating.

For the first frame, the model landmarks, which have been obtained as explained in Sect. 2, are also used as landmarks for that frame. For the subsequent frames, once the recognition procedure is finished, the centroids of the facial features of interest are calculated. Also, the centroids of the pre-defined regions in the model are calculated. Then, the affine transformation that best maps the model set of centroids to that of the considered frame is estimated and applied through the following formula [5]:

$$\overrightarrow{q} = \alpha(A\overrightarrow{s} + \overrightarrow{b}) \tag{7}$$

where A corresponds to a 2×2 non-singular matrix representing the sought transformation, α is any scalar value, and $\overrightarrow{q}$, $\overrightarrow{s}$ are the centroid-coordinate vectors for the frame and model respectively.

This affine transformation allows us not only to update the input face landmarks to be applied to the following frame in the process, but also to project the model image onto the segmented and recognized target frame, conveying a visual assessment of the matching process.

Possible Extensions. Although this change in the methodology already makes it more robust for the application in video sequences, our ongoing research aims at making better use of the possibly redundant information present in distinct frames.

One possible approach is to insert *temporal edges* to the set E of an ARG G. These edges would represent the transitions and relations among vertices of consecutive frames in the sequence. Through this, it would be possible to recalculate both vertex and edge attributes and a model image could be no longer needed for the recognition. Also, features which were not present in the first frame could be added to the recognition process on-the-fly.

Furthermore, the results obtained in the graph matching procedure for, say, frame i could be reused as the initial solution for the matching step in frame $i+1$, thus reducing the tree expansion and taking into account the smooth changes presented in frame transitions. This generalization belongs to our ongoing research.

5 Results and Conclusion

In this section we show some of the first results obtained from the application of the new steps introduced in the previous section. For the tests, different video sequences were considered, such as sequences of male and female faces with static or changing background. All sequences presented considerable changes in the face (e.g. smiles, head movement, blinking) throughout time.

Figure 2 depicts the results obtained for the frame-to-frame projection of the model-mask onto the corresponding target frames. The video sequence was composed of 96 color frames of size 512 × 512 which have been converted to grey-level images for the purpose of the algorithm. As it can be seen, the model mask is successfully matched up to the face, thus allowing it to be tracked along the video sequence. The facial features defined by the mask are approximately matched up to their correspondents in the image, though some mismatched regions (mouth in the last image) may be noted due to the global nature of the affine transform and to differences in the facial expressions among the frames.

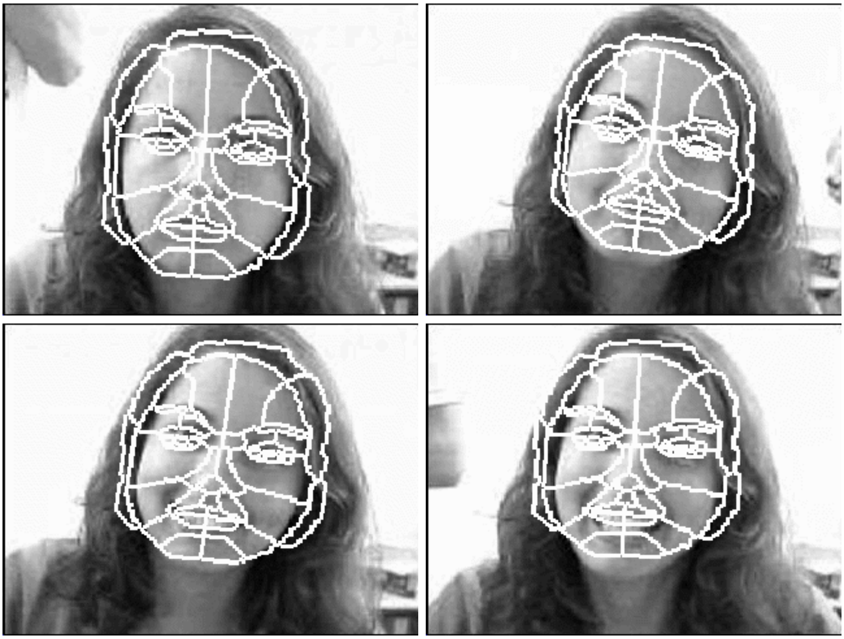

Fig. 2. Model masks superimposed on successive target frames using the recognized facial features.

In terms of the results obtained by the advances proposed in this paper, i.e, the landmark updating and its assessment through the projection of the model mask onto the input face, it can be seen that the mask projection follows the head movements in a plausible manner. Also, most facial features which may be of interest are correctly tracked (e.g.: eyebrows, nostrils, nose, lips), showing that the recognition process and the landmark updating can be effective and provide encouraging results.

Nevertheless, certain refinements in the technique are still called for, especially when a considerable sudden change is present between frames, or when unknown facial features, i.e. those which were not present in the model, appear throughout the sequence. In such cases, the unknown facial features will be necessarily mapped to one of the classified facial features, which might lead to results such as the one seen in the frames of Figure 2 where a smile occurs.

Thus, in this paper we have proposed a first approach towards the generalization of the methodology presented in [3]. The first results have shown that it is possible to reflect the changes in facial features in each frame that occurs throughout time using appropriate affine transformations. Although the introduced steps have provided encouraging results, the other possibilities mentioned in Sect. 4 are being considered in our ongoing work.

Acknowledgments. This work was partially funded by a CAPES/COFECUB grant. A. Graciano is grateful to CNPq (132562/2003-0). R. Cesar is grateful to FAPESP (99/12765-2) and to CNPq (300722/98-2).

References

[1] O. Colliot, A.V. Tuzikov, R.M. Cesar Jr., and I. Bloch. Approximate reflectional symmetries of fuzzy objects with an application in model-based object recognition. *Fuzzy Sets and Systems*, 2003. In Press.

[2] D. Fontaine and N. Ramaux. An approach by graphs for the recognition of temporal scenarios. *IEEE Transactions on Systems, Man, and Cybernetics - Part B: Cybernetics*, 28(3):387–403, June 1998.

[3] R.M. Cesar Jr., E. Bengoetxea, and I.Bloch. Inexact graph matching using stochastic optimization techniques for facial feature recognition. In *16th International Conference on Pattern Recognition*, volume 2, pages 465–468, August 2002.

[4] R.M. Cesar Jr. and I. Bloch. First results on facial feature segmentation and recognition using graph homomorphisms. In *Proc. VI Simpósio Ibero-Americano de Reconhecimento de Padrões*, pages 95–99, Florianópolis, Brazil, 2001.

[5] R.M. Cesar Jr. and L. da F. Costa. *Shape Analysis and Classification – Theory and Practice*. CRC Press, 1 edition, 2001.

[6] V. Kruger and G. Sommer. Affine real-time face tracking using a wavelet network. In *Proc. of ICVV'99 Workshop Recognition, Analysis, and Tracking of Faces and Gestures in Real-Time Systems*, pages 141–148, Corfu, Greece, 1999.

[7] A. Perchant and I. Bloch. Fuzzy morphisms between graphs. *Fuzzy Sets and Systems*, 128(2):149–168, 2002.

[8] M. Shah and R. Jain, editors. *Motion-Based Recognition*. Computational Imaging and Vision. Kluwer Academic Publishers, 1997.

[9] L.B. Shams, M.J. Brady, and S. Schaal. Graph matching vs mutual information maximization for object detection. *Neural Networks*, 14:345–354, 2001.

[10] R.C. Wilson and E.R. Hancock. A Bayesian compatibility model for graph matching. *Pattern Recognition Letters*, 17(3):263–276, 1996.