Defining Search Areas to Localize Limbs in Body Motion Analysis

Thomas Fourès and Philippe Joly

Institut de Recherche en Informatique de Toulouse
Université Paul Sabatier
118 Route de Narbonne
31062 Toulouse Cedex 4
France
{Thomas.Foures, Philippe.Joly}@irit.fr
WWW home page: http://www.irit.fr/recherches/SAMOVA

Abstract. This paper deals with the use of a model dedicated to human motion analysis in a video. This model has the particularity to be able to adapt itself to the current resolution or the required level of precision through possible decompositions into several hierarchical levels. The first level of the model has been described in previous works: it is region-based and the matching process between the model and the current picture is performed by the comparison of the extracted subject shape and a graphical representation of the model consisting in a set of ribbons. To proceed to this comparison, a chamfer matching algorithm is applied on those regions. Until now, the correspondence problem was treated in an independent way for each element of the model in a search area, one for each limb. No physical constraints were applied while positioning the different ribbons, as no temporal information has been taken into account. We present in this paper how we intend to introduce all those parameters in the definition of the different search areas according to positions obtained in the previous frames, distance with neighbor ribbons, and quality of previous matching.

1 Introduction

By the expansion of applications such as surveillance systems, user interface, or in more cultural domains, sport or dance motions analysis, study of human motion in video sequences has become a domain in wide expansion. The main objective is to be able to identify some predefined gestures in the motion description provided by this method. This identification step will depend on the reliability of results produced by the analysis step, and the ability to recognize on this description a same gesture shot from different view angles. There are two major applications in the field of video content indexing for these works: we intend first to be able to identify segments in sport events where a specific gesture occurs; we also want to apply this tool to video surveillance in order to automatically detect some given human behavior in order to launch an alarm on purpose.

All systems differ considering the means involved to achieve this analysis. They depend directly from the application itself, its context, and some requirements imposed by user needs. The way human body is modeled is one of the most critical parts, and in our case, we will have to deal in real time with a video flow of low quality. Therefore, the model we propose can adapt itself to various video qualities and can be used to produce motion descriptions at different levels of details.

Different kind of human models exist (as for example H-Anim from MPEG-4 [5]), meanwhile they are essentially derived for synthesis purposes and then are not in total adequacy for an automatic analysis tool.

In this paper, we propose a multi-level model for the human body and we discuss about the way to impose physical constraints on it in order to reinforce results. The principle is to define only one model for any kind of document. It can adapt itself to the current resolution or precision level required by a user, avoiding by this way useless computation and prevent from false detection as it may happen when the model is defined with more accuracy than what it is really possible to extract from the video flow. To achieve this, the model is composed of ribbons, each of them being decomposable in sub-ribbons, implying a descent into the hierarchical levels. Thus we can only use the level corresponding to the document resolution or to specific needs. The matching process between model and frames extracted from a video consists in a chamfer matching algorithm between model components and a distance map. This map is obtained from a distance transform applied on the subject image, but this step has to be preceded by a decomposition of the image in search areas for each component of the model.

Defining a model in this hierarchical way has two advantages. First, dealing with real time processing becomes possible as a result even at a coarse degree can always be produced. The obtained precision will depend on the time allowed to perform computing. The more time being available the more results are accurate. The second advantage is the possibility to adapt the model to application needs. Users can specify the required degree of precision and the appropriate model decomposition is used in consequence, avoiding useless computation cost.
The definition of search areas is a critical point in the process. At the beginning, they are estimated according to the most likely position of the limbs in the image. As a first match with the model has been performed, we intend to refine the location of those search areas by taking into account matching quality, and introducing physical constraints on the positioning of the different elements in the image.

2 Related Work

This section has for objective to give a rapid overview of related works, and justify in the same time the proposed approach. As mentioned before, human models are often inspired from image synthesis. In this domain, the model developed by MPEG-4 [5] is composed of "sticks" linked by "joints". But that kind of modeling presents some limitations because it is based on the idea that

articulations are centers of rotations. It has not been designed for analysis purposes, which makes its use difficult in this context where limitations on possible subject postures must be imposed. It seems difficult to specify constraints on for example, rotation angles or even limbs velocity. Nevertheless, we can find a use of that kind of model in [7] where a matching of the skeleton obtained from a subject shape is performed. Another kind of modeling has been proposed in order to derive directly the model from original images themselves. It is the case with statistical approaches. We can distinguish among them the use of spatial and color distribution of pixels in order to realize the image segmentation [9]. But in that case, motion description depends on environment conditions and can often provide a too rough description, difficult to employ in more general cases as we can encounter in multimedia information retrieval. The second kind of statistical approaches consists in models which are "deformable" since they have the particularity to adapt to the subject images ([3], [8]). By using point distribution models, the model shape becomes "active". But this approach requires a training step for the matching process to provide good results [4].

No modeling proposition integrates a multilevel resolution of the description. Furthermore, most of those approaches require some strict conditions on the way the analyzed scene has to be shot (conditions on lightening, camera centering, etc). This reduces their potential application for generic and real time video content analysis. Even if our proposition do still not take into account 3D information, truncated shapes, camera motion, or multiple bodies, we think that these points can be integrated in further works in a compliant way. Whatever, (in the generated description) the reliability of results, which allows trusting only in pieces of information of good quality, is already taken into account. This point is typical from an indexing approach, which is a new way to address human motion analysis.

3 Previous Work

3.1 The Hierarchical Model

As mentioned before, the proposed model is defined in a hierarchical way: it is composed of many levels, each of them corresponding to a level of details which best feat the video resolution. The principle is to perform a first processing step by using the first level of the model providing the roughest results [6]. Then, these results can be refined by the application of the second level which provides a sharper description of the human body. For elements of the model where results are better with a more precise description, the next level can be applied, and so on until application of a new level providing results which are not more significant that before. The final level will be the one which feat the best the resolution of the video, the level of detail which is really extractible.

Possible applications can be set in three different groups. The first case where no knowledge about the studied document is supplied and no real time processing is required. Then a descent into the hierarchical levels is performed and best

matching selected. This approach yields when a temporal constraint is introduced. In that case, the accuracy of the description is limited by computation time, but as we said, the proposed model is by its conception compliant with real time processing, and then, a result (even coarse in comparison with what is really extractable from the video) can be produced. This property can not be offered by a non evolutive model where all the components coordinates are required to define a subject position. The third possible case is when the user whish to specify the accuracy level of the application. For example, in the same video, a sharp description of the motion of the arms and a coarser of the legs may be required. Then, the model can be composed of high level components for the top of the body, and low level ones for the rest of the limbs. In that last case, an adaptation to specific needs is achieved.

Considering those orientations, we are able to give a definition of our model which is graphically based on regions, and this feature is essential for the forthcoming process which relies on this characteristic. Region-based means that the surface covered by one element of the model has to match with the corresponding subject limb in the image, the model being composed of a set of ribbons, each of them is associated with one part of the human body. The evolutive aspect of the proposed model stays in the fact that those ribbons can be decomposed in sub-ribbons, and by this way the model has the possibility to be adapted to the required level of precision. Defining our model by simple elements as ribbons and sub-ribbons instead of more developed features as edges for example, has the advantage to require a simple description. Indeed, an element of the model can be defined by two parameters, its length and width. The localization in the image is realized by also two parameters: the coordinates of a control point and an angle value which corresponds to the orientation of the concerned segment. Thus, an element of the model is totally defined by a couple of parameters to generate it and a couple of parameters to localize it in the image. Obviously, increasing the resolution (implying a descent in the model hierarchy) will raise the number of elements and in the same way, the number of parameters required to define the posture of the studied subject. Based on this idea of multi-level model, we propose 3 different descriptions (see Fig. 1). The first one is quite basic since it is composed by only 5 ribbons, one for each limb and one for the head-torso set. This representation is rough and most limbs configurations will not be able to be described in a precise way. However, a first description of the posture can be given by this kind of modeling and may be sufficient for some applications which do not require a good precision. On the second level, all those 5 ribbons split up at the joints area and allow detection of a more important number of limbs configurations such as, for example bending. Finally, a third level has been proposed. By using this one, it is possible to distinguish motion of hands and feet by splitting up the model into components corresponding to the forearms and the down part of the leg. The choice of this decomposition into three different levels has been made according to what seemed to be feasible from different image resolutions. Of course, many other configurations based on this idea of a hierarchical model are possible.

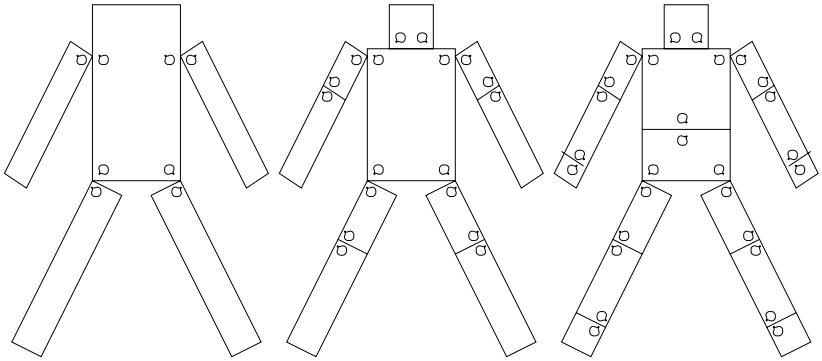

Fig. 1. Graphical rendering of the proposed model. From left to right: the three hierarchical levels.

3.2 Model/Subject Matching

Our model being defined, we have to describe methods used to match the different elements which compose the model and extracted video features. The subject posture is indicated by the obtained positions for those elements. The first step of our process is to match the first level of the hierarchical decomposition which, in all cases, provides first information about the limbs localization and can be employed during next steps when more refined levels of the model are involved. By this way, using a coarse definition for the model specifies the search areas and then provides a reduction of the computational cost for the application of the following levels of the model.

Before the model/subject matching step, the first operation consists in the extraction of features from the video flow which are compliant with the region-based model. This is realized by preprocessing a background subtraction followed by morphological operations in order to obtain an image of the subject silhouette. Then boundary boxes are created, giving a first reduction of the search field by keeping only regions of interest, and through their length/width ratio define the size of elements composing the model.

The problem is now to determine the correspondence between those silhouette images and the different elements of the model. We use for that a chamfer matching algorithm. Its principles are exposed in ([1], [2]). The method consists in the construction of a distance map computed with the chamfer distance on the binary image of the shape where the searched feature is located. Then a distance between the image of the feature and this map is evaluated and allows to know if the proposed position of the model is acceptable or not and, in the negative case, to estimate the difference value. The distance transformation, i.e. the conversion of a binary image into a distance map, is processed by using the 3-4 Distance Transform defined by Borgefors [2] allowing a good approximation of the chamfer in only two passes over the image.

The distance map being created, each element of the model has to be used as a mask over this map and the root mean square of the distance values located under the mask is computed. The r.m.s. is the difference measure presented by Borgefors as being the one providing the less false minima among other existing average measures. The search of all the components positions is performed in a sequential way: for several successive values of the model parameters, only the ones corresponding to positions providing the minimum r.m.s. value are selected. Those parameters are spatial coordinates $(x; y)$ of the segments and the value of the angle with one of the two axes. Thus, the search for minimum r.m.s. has to take into account all possible orientations of segments in addition to their spatial coordinates. The implemented algorithm (see Fig. 2) begins with a predefined initial position and orientation. Then, a matching step using only translations is processed, and from the new determined position, a rotation of $\pm\frac{\pi}{4}$ around the center of the intersection area between the component image and the subject silhouette is applied. For those angle values, a new matching by translations is performed. We select the position with the lower r.m.s. value among the initial and the two new ones. In order to refine results, this process is reiterated by using for the rotation an angle equal to the preceding one divided by 2. This time, the comparison is processed between the previously computed position and the two new ones coming from this last step, and only the best one is kept. This operation is repeated 6 times, corresponding to a rotation angle value equal to $\pm\frac{\pi}{128}$, which provides a low error at the pixel scale. The choice of $\frac{\pi}{4}$ for the first rotation is due to the symmetry of the model components which ensure that even by limiting the angle to this value all possible orientations will be explored. The algorithm convergence is based on the fact that matching error is minimal when a model segment is globally oriented in the same direction that the searched subject limb. Matching by only translations is a process that reduces effects coming from the potential presence in the neighborhood of pixels which do not belong to the area of interest. The general implemented algorithm is described in Fig. 2.

The chamfer matching algorithm has the disadvantage to lead to potential false detections when the initial position is too far from the optimal one. To avoid this kind of situation, we define a search area for each element composing the model (this step corresponds to the box labeled "Image cutting" in Fig. 2). As no a priori information about subject posture is available, the current image is cut out according to the most probable location of subject limbs. The definition of those areas presents the advantage to reduce the search field, and then, in addition of avoiding local minima, provides a gain in terms of computational cost. On the other hand, if the searched limb is not located in the affected search area, it becomes impossible to localize it. Thus, the way areas are defined is a critical point of the system.

3.3 Experimental Results With Non-Evolutive Search Areas

In this section, a few matching results using only the first level of the model and non-evolutive search areas are presented. The search areas are called "non-

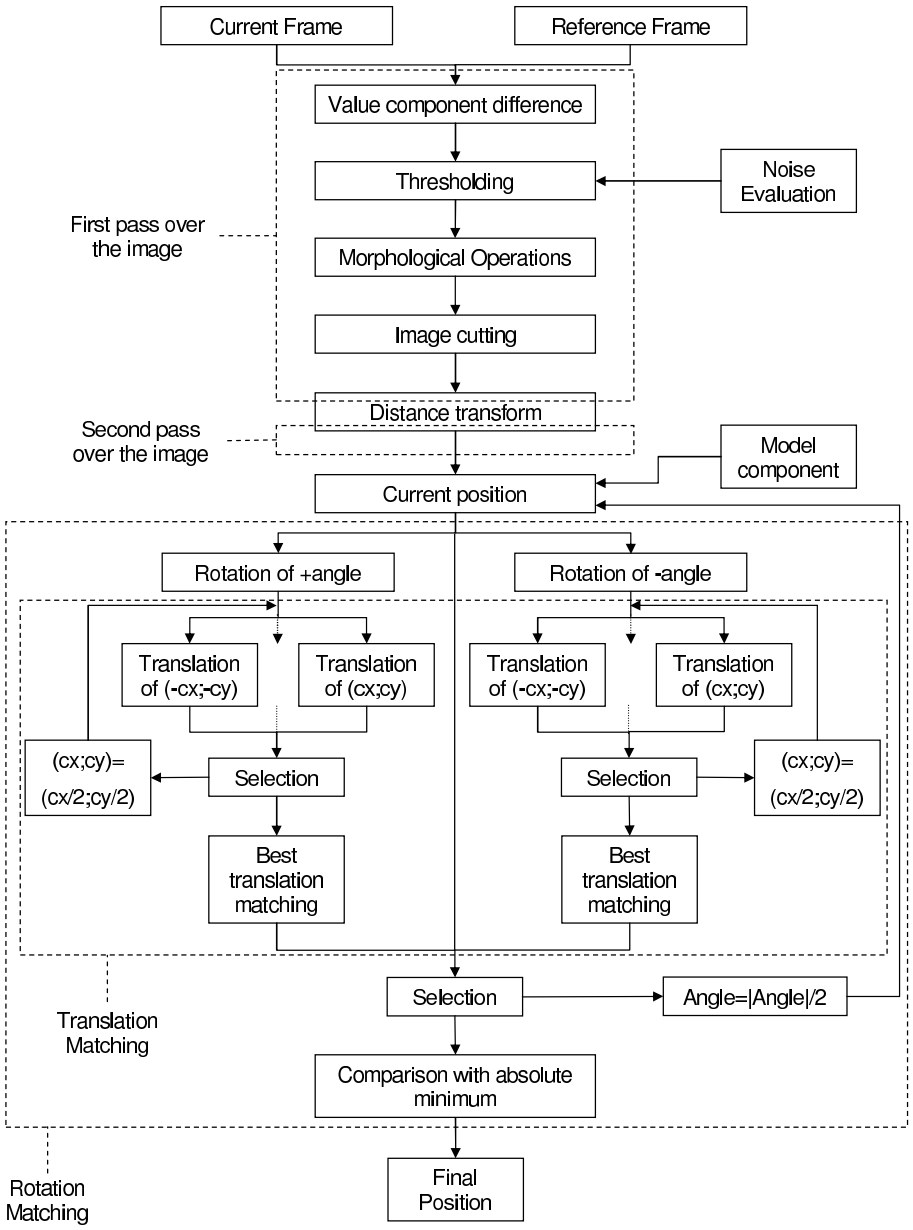

Fig. 2. The general algorithm from a frame to model coordinates. In two passes over the image, all preprocessing steps and a distance transform can be computed. They are followed by an algorithm aiming at matching with translations and with rotations. Translation process is performed until the translation factor (cx;cy) is different from zero, and rotation process is as for him repeated until the angle value is greater than $\frac{\pi}{128}$.

evolutive" because their definition has been made according to the most likely localization of the limbs in the image at the beginning of a sequence and do not take into account some information coming from previous matching steps (see Fig. 3 for initial image cutting details).

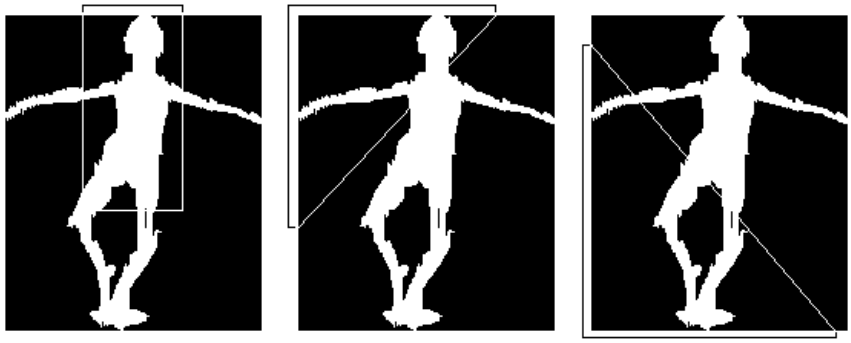

Fig. 3. Subject silhouette obtained after preprocessing. Initial search areas for respectively torso, right arm and right leg are delimited by a white line (within the image) and black ones (outside). Fields corresponding to left arm and leg are the symmetrical ones.

Even if the first level of the model offers some restricted possibilities, matching results are not so far from the real subject posture. Some differences appear when for example, some pixels that do not belong to the concerned limb are located in the search area, or when there are some "holes" in the subject silhouette that preprocessing is not able to fill. Obviously, some limitations about the limbs configuration of the subject are imposed by the high degree of this first level, and then some postures can not be analyzed. In addition to that, many problems of analyzing a human motion come from the choice of a 2D model and of only one camera. Some postures where depth of field is important are difficult to describe because the 3D information is missing. Only data about motion from a global point of view could allow a description of subject postures in that kind of situations.

In order to evaluate the quality of matching, a measure has been proposed. Its objective is to provide an indication about the validity of a positioning for each element that composes the model. The used formula is:

$$Q_1 = 2 * \frac{Nbr\ of\ pix.\ in\ the\ area\ of\ real\ image \subset model\ component\ matched}{Nbr\ of\ pix.\ model\ component} - 1\,.$$

The obtained value for Q_1 is between -1 and +1; -1 is a non significant matching that can not be employed, and +1 on the contrary a reliable one, useful for next pictures study. It is important to precise that this measure provides an information on only the surface covered by a model element, and do not take into account the fact that the detected limb is the searched one or not. This

knowledge can only be brought by specifying constraints on the distance between different model components, and evaluating the validity of determined posture.

On the example of silhouette given by Fig. 1, we can notice that because of the thresholding realized in the preprocessing step, some part of the subject have not been well highlighted and some "holes" appear within the silhouette. However, those pixels are not taken into account to compute a quality of the matching, whereas they certainly belong to the subject limb. To solve this problem and provide a better evaluation of the matching, we propose to compute a coefficient Q_2 from the distance map:

$$Q_2 = f\ (\frac{Values\ of\ distance\ map \subset model\ component\ matched}{255 * Nbr\ of\ pix.\ model\ component})$$

where f is a function intended to spread out values between -1 and +1, providing by this way a better results interpretation. This function is experimentally determined, in order to lead to values which are in agreement with real subject posture.
As Q_1, Q_2 takes its possible values between -1 and +1. Normalization is processed by considering that 255 is the maximum value fixed for an element of the map. This limitation has been determined by taking into account the size of the bounding boxes and this value can be seldom reached. In what follows we will get use of the letter Q to describe the coefficient of matching quality which is a function of the coefficients Q_1 and Q_2.

Table 1. Q_1 and Q_2 values obtained for the images given on Fig. 4, each column corresponds to an element of the model for each one of the two coefficients.

	Q_1					Q_2				
	T	RA	LA	RL	LL	T	RA	LA	RL	LL
Pic.1	0.5255	0.0240	0.2678	0.0388	0.3682	0.8140	0.5040	0.7860	0.5450	0.7900
Pic.2	0.5719	0.0135	0.1329	0.2409	0.3117	0.8190	0.7000	0.6960	0.6130	0.7410
Pic.3	0.1673	0.4058	-0.2433	0.1284	-0.3517	0.5890	0.6820	-0.5560	0.5300	0.2780
Pic.4	0.3633	-0.0049	0.0323	-0.0867	-0.2624	0.7000	0.6600	0.6730	-0.2280	-0.7150
Pic.5	0.2807	-0.7159	0.2901	0.1447	-0.2340	0.6404	-1	0.6937	0.6580	0.2638
Pic.6	0.1704	0.4915	-0.4145	-0.0421	-0.7141	0.5607	0.8480	-0.1020	0.1312	-1

As we can see on the graphical rendering of evaluated postures (see Fig. 4), the proposed algorithm produces quite good results when the objective is to match a model element with pixels located in a precise area. But this operation will provide the detection of given limb only if search areas have been correctly defined. Thus, the way those areas are defined is determining, as a limb can not be detected if it does not belong to its search field. Obviously, this implies that if we do not have a priori knowledge about the performed motion, an initialization step during which subject limbs have to cross their respective search areas is necessary to be able to begin a tracking process. To achieve this tracking step,

an evolution of the areas according to previous matching, its quality, and physical constraints must be proposed.

4 Search Areas Redefinition

At this point of development, the goal is to incorporate into the system some knowledge coming from parameters likely to provide some relevant information. By today, the matching of an element of the model was performed independently from the others. For example, no information about torso positioning was used to lead the arms matching, or any other limb. This consideration highlights the necessity of modeling the physical constraints inherent to the human body. The evaluation of matching quality has also to be taken into account by the algorithm. Indeed, the surface we have to cover for a search do not have to be the same if previous matching could be considered as excellent, or on the contrary was not good at all. Of course, these two parameters (physical constraints and matching quality) are not totally unrelated, and their intended usage must be a good compromise between their possible contributions. A possible solution to incorporate them into the matching process could have been to modify directly the distance map according to them, giving a new value to some map elements in order to give some orientations to the search. But obviously it seems that the generated computing complexity makes this possibility an inappropriate solution. Consequently, we propose to incorporate the constraints (of quality and also of physical type) to the definition of search areas which is a principal part of the system on which relies most of the potential efficiency.

4.1 Application of Physical Constraints

We intend to use the fact that, for example, the element corresponding to the torso can not be at a given distance from the ones representing arms and legs. By this way, a model of tensions involved in the human body can be realized. It is important to precise that we do not want to force the localization of an element precisely by the torso side (for example), but rather to lead the next search in a more probable direction in order to save computation time and avoid wrong matching. We propose to realize this operation of bringing towards model elements that are supposed to be joined by redefining search areas according to previously determined limbs localization.

The goal is, once a search area has been determined, to move this field towards the adjacent limb. This operation can be processed in many ways: two adjacent segments can be brought to get closer one to the other, or a segment can be considered as static, and only one has to move to get closer. The quality coefficient Q can be an indication to take the decision of which area can be considered as static but only in very restricted proportions because, as we mentioned before, it just provides an information in terms of covered surface but not about the limb detected indeed. In a first step, in order to evaluate those principles on a simple case, we suppose that the segment representing torso is

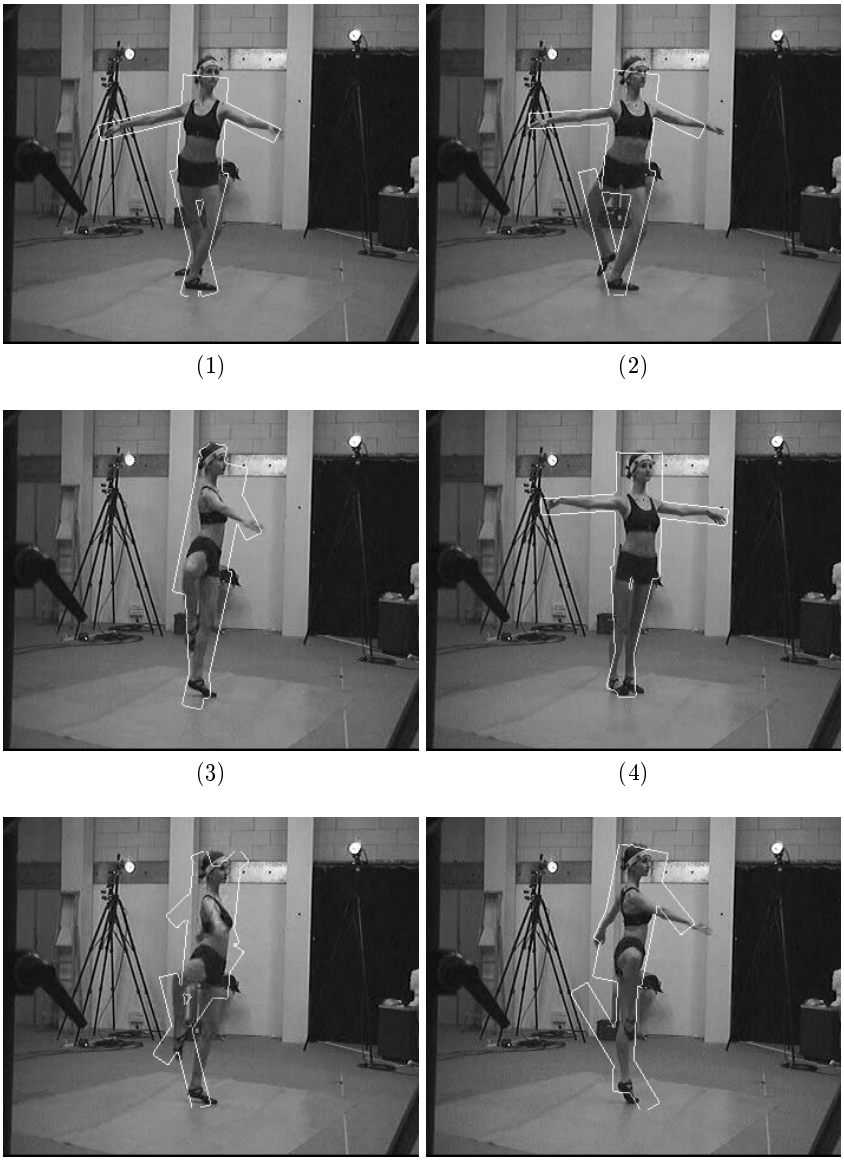

Fig. 4. Frames extracted from a video. Results of the matching process are represented by white lines around the subject.

the one the more likely to be effectively matched with the searched limb. This assumption is made according to the fact that the torso is the only one limb we are almost sure to have detected without too much ambiguity when a subject is present, even if the localization is not precise in a first step and need to be refined. Thus, this model element can be chosen as a reference, and we have to move the search areas of all the adjacent components towards it.

The figure 5 is an illustration of those principles in a simple case. This is the result of a first matching. As we said, we suppose that the evaluated position for the head-torso set can be considered as quite good (even if it will also have to evolve in order to get more accurate). The search area of the arm for example has to be redefined, first in terms of surface (this problem being tackled in next section), and then in terms of position. To solve this last point, the information concerning torso matching in the previous frame is used to transfer area coordinates towards the corresponding part of the torso. A translation moves the search area towards an intermediate position between the original one and the torso extremity. We proceed by an interpolation and not a complete translation because the torso localization is not accurate enough (as for the other limbs) and may evolve in the time. Proceeding to an interpolation should realize a smoothing effect on the segment displacement, avoiding on the same time some possible oscillations. By this way, the search field evolves towards an area of the image in which searched limb is supposed to be located. This redefinition is based on a part of the result obtained from the previous matching. Thus, we intend to provide a valid redirection for the search fields definition. To determine in which proportions an area should be moved, we propose to use the matching quality given by the coefficient Q. The proposed formula for the translation factor $T(x; y)$ is:

$$T(x; y) = \frac{Q_A}{Q_T} \left(R_A(x; y) - P_T(x; y)\right)$$

with:

- Q_A, Q_T coefficients of matching quality for the concerned limb (here the arm) and the torso
- R_A initial position of the limb search area
- P_T position of the torso segment.

Torso being considered as a reference for other matching operations, its search area has nevertheless to be redefined in order to provide an accurate positioning. We intend to realize this operation by a balanced interpolation from positions of the four other search areas. Only a translation of the search field is performed, the orientation obtained in the previous frame for the element is preserved as orientation for its redefinition. This angle represents a general orientation of pixels within the area and provides some information that should be used in the areas redefinition process. We can notice that restricting the distance between joints of adjacent areas and redefining the different search fields as explained before limit, in most cases, the surface which may be covered by different areas, excepted of course when a limb performs a motion leading to pass behind another one, where in that case recovers are inevitable.

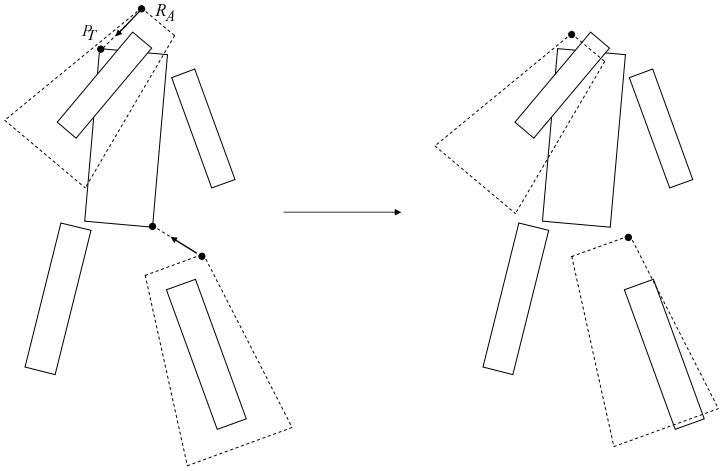

Fig. 5. Application of physical constraints on search areas definition. From an obtained posture (on the left), search areas are first defined and then relocated by using control points. New areas after this operation are illustrated on the right.

4.2 Application of the Quality Matching Coefficient

In the previous section, we have discussed about the localization of a search field around the determined area of an element of the model, however the surface covered by this area has still to be defined. To achieve this, we propose to use the coefficient Q representing the quality of a matching to evaluate well-proportioned dimensions for the search areas. Indeed, the less pixels are presents in the current search field (that means a low value for Q), the more this field has to grow in the next frame in order to extend the possibilities to find the limb. On the contrary, if Q is high enough, then many pixels belong to the search area and in that case, the current matching can be considered as quite reliable. In addition to that, if the obtained position for the different elements of the model is conform to a human body configuration (actually compliant to physical constraints), we are allowed to suppose that the real limb has been detected. Then, to process the next frame, it is not necessary to search in a wide area. We can restrict the search to the near neighborhood.

4.3 Proposed Parameters

Two aspects have to be taken into account to redefine a search area: physical constraints coming from the human body constitution and up to what point a previous matching must influence the next one. We have to propose a mean to introduce those constraints into the parametrical definition of a search area. This one can be defined from a single point and an angular value (see Fig 6). We use an angular sector because this kind of parameter seems to be more adapted to

motions of model elements, and allows more accurate descriptions. The position of the point O and the value α will directly depend on the physical constraints and the quality of the previous matching. We propose to set O on the line issued from the point located in the middle of a model element width (the part of the element nearest to the torso), keeping the segment orientation computed in the previous step. Another possibility could have been to fix O on the joint with the torso (or near to this point). In that case, the only parameter defining the surface area would be α. This last solution provides certainly a simplification of the area computation by reducing the parameters number, but it also seems that in most of the cases, the final surface obtained by this way will be bigger in terms of covered pixels and then will require a higher computational time. Furthermore, it would not be able to provide a sharper definition as the proposed solution. At last, the non-exploitation of the previous limb orientation represents a certain loss of relevant information.

The distance ρ between the point M and O (see Fig 6) is function of the quality of the previous matching: the weaker Q, the higher ρ, in order to have maximum of possibilities to describe the search area precisely for each different cases; ρ depends on Q. From O, the angular value α will partially fix the surface covered by the search area. This parameter has also to rely on Q by evolving higher when this one gets lower: if the matching was good, we do not have to define a wide area, and on the contrary, if the matching was not good enough, we have to extend this surface to recover pixels belonging to the concerned limb. Possible values for α is a function of Q. To close the search area, we still have to provide a value for the height h of the area. This is another parameter used to define the search area. For the moment, only the quality of matching has been employed to describe the search field in terms of surface. Physical constraints of the human body will be applied while localizing this search area in the image space as seen in previous section. By reducing the distance between the area and the obtained position of the model element corresponding to the torso, we intend to focus the search in this direction and then on the most likely part of the image. Of course, this implies that torso positioning is seen as a reference for the other limbs, but it has also to evolve itself by moving its own search area as a rectangular surface around it, this area having the same orientation in the previous frame. We do not use an angular sector to define this search field. Indeed, the motion of the torso do not necessarily requires a very accurate description for the search area, because of the size of this limb in the bounding box. The temporal evolution of the torso orientation is generally not in the same proportions that the one of others limbs (as arm for example), and then do not need an expensive computation time. On the other hand, the dimensions of this rectangular area have to evolve with the matching quality. The localization of this search field has to be processed as described in the previous section.

In the next levels of the hierarchical model providing a more accurate description of adopted postures, the different search areas will be redefined by parameters described previously: a single point and an angular value. This time, we propose to take as the reference limb the one at the immediate upper level in

the hierarchy. For example, the segment representing the forearm will have an area which depends on the position of other elements composing the full arm. Another kind of hierarchy (this time between the limbs) is involved. However, this dependency in the areas redefinition should not be a constraint preventing from the matching process when a limb of a higher level has not been correctly detected.

We mentioned that in a first step, the orientation of the search area will be the same as the one of the concerned limb in the previous frame. This assumption comes from the fact that between two consecutive frames, the motion can not be fast enough to produce a significant change in the orientation. However, this method to determine the orientation can not be applied when only key frames are processed, because high variations may happen. This time, motion dynamic must be evaluated in order to estimate the new orientation of the search area.

In order to define the search field, four points (R_1, R_2, R_3, R_4) delimiting this area are required. Those point coordinates in the image plan will be directly computed from parameters ρ, α and h described above. They are obtained by performing two changes of reference marks: first, O is defined by its polar coordinates in the reference where the point P is the center. Then, coordinates of R_1, R_2, R_3 and R_4 in the polar reference mark of center O are computed. At last, we are able to give the coordinates of the four points in the Cartesian image reference. Obtained formulas are given below:

$$
\begin{aligned}
R_1 &: (R_{1x}\,;\, R_{1y}) = (X_O + \rho_1 sin\theta_1 + P_x + T_x\,;\, Y_O + \rho_1 cos\theta_1 + P_y + T_y)\\
R_2 &: (R_{2x}\,;\, R_{2y}) = (X_O + \rho_2 sin\theta_1 + P_x + T_x\,;\, Y_O + \rho_2 cos\theta_1 + P_y + T_y)\\
R_3 &: (R_{3x}\,;\, R_{3y}) = (X_O + \rho_2 sin\theta_2 + P_x + T_x\,;\, Y_O + \rho_2 cos\theta_2 + P_y + T_y)\\
R_4 &: (R_{4x}\,;\, R_{4y}) = (X_O + \rho_1 sin\theta_2 + P_x + T_x\,;\, Y_O + \rho_1 cos\theta_2 + P_y + T_y)
\end{aligned}
$$

with:

- $(X_O\,;\, Y_O) = (\frac{\rho}{sin\delta}\, sin(\beta + \frac{\pi}{2} + \delta)\,;\, \frac{\rho}{sin\delta}\, cos(\beta + \frac{\pi}{2} + \delta))$ Cartesian coordinates of O in reference mark of center P
- $\rho_1 = \frac{\rho - h}{cos\frac{\alpha}{2}}$ is the modulus of R_1 and R_4 is reference mark of center O
- $\theta_1 = \beta - \frac{\alpha}{2}$ is the argument of R_1 and R_2 in the same reference mark
- $\rho_2 = \frac{\rho + h + L}{cos\frac{\alpha}{2}}$ corresponds to modulus of R_2 and R_3
- $\theta_2 = \beta + \frac{\alpha}{2}$ corresponds to the argument of R_3 and R_4.

In addition to this, we have to precise that:

- L is the length of the concerned segment of the model
- δ is the angle between (OP) and the element (see figure) and is equal to $atan(\frac{\rho}{\frac{l}{2}})$
- β is the orientation obtained by the matching process
- $(P_x\,;\, P_y)$ are the coordinates of point P in the image
- $(T_x\,;\, T_y)$ is the translation factor computed according to concepts exposed in the previous section
- l is the element width.

The coordinates of each extremity of a search area can be computed from three parameters which are ρ, α and h. These parameters are function of the matching quality coefficient Q obtained at a previous processing step.

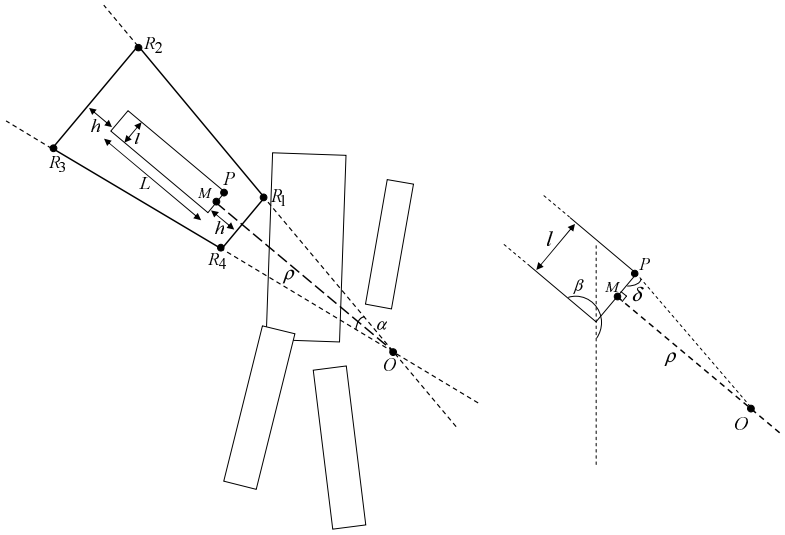

Fig. 6. Redefinition of the search field corresponding to the arm. Coordinates of each extremity of the area are computed from the position of the model element. On the right an enlargement highlights angles used to perform computations.

4.4 Future Works

Future works will mainly determine in which proportion each parameter (ρ, α and h) has to be used in order to evaluate for all the different cases the best size for the new search areas. As mentioned before, the other point of work is to establish the relation with the previous matching. Another aspect is the temporal evolution of the parameter values. Some oscillations are likely to happen, and limitations on possible results should be applied in order to avoid those effects, as an interpolation has been introduced in the translation process of the search areas with the same objectives. Introduction of other conditions in the definition of the search fields may also be considered. For example, an evolution taking into account the other areas redefinitions according the most likely ones, creating by this way a kind of evolutive hierarchy between the different limbs, is a possible way to be explored. Obviously, this will imply to take a decision to order segment positions, as we have done by choosing the torso as the reference for first matchings.

5 Conclusion

We have proposed in this paper methods dedicated to incorporate some temporal and spatial information to the refining of limbs localization in a system performing a subject/model matching. This system already gave first results of matching at the coarsest level of a hierarchical model. Our goal was to propose

means to provide a more accurate positioning of the different elements. In order to achieve this, we have first introduced a direct application of the physical constraints inherent to the human body through a modification of the localization of the next search fields, these ones being essential to improve the results quality. Torso is considered as a reference limb for the other elements during this operation. The second step has been to define those search areas in terms of covered surface by taking into account the previous matching quality. This has been realized by using an angular sector generating a surface according to the area to be explored. With these two processes, new positions and sizes of the search fields can be defined. The next step of our works will be to provide an experimental validation of all the exposed principles and to define in which proportion each parameter occurring in new search areas computation has to participate to the area definition.

References

1. H.G. Barrow, J.M. Tenenbaum, R.C. Bolles, and H.C. Wolf. Parametric correspondence and chamfer matching: two new techniques for image matching. In *Proc. 5th Int. Joint Conf. Artificial Intelligence*, pages 659–663, Cambridge, MA, 1977.
2. G. Borgefors. Hierarchical chamfer matching: a parametric edge matching algorithm. *IEEE Trans. Pattern Anal. Mach. Intell.*, 10(6):849–865, November 1988.
3. T. Cootes, A. Hill, C. Taylor, and J. Haslam. The use of active shape models for locating structures in medical images. In *IPMI*, pages 33–47, 1993.
4. T. Cootes, C. Taylor, D. Cooper, and J. Graham. Training models shape from sets of exemples. In *Proc. of British Machine Vision Conference*, pages 9–18, September 1992.
5. T. Ebrahimi and F. Pereira. *The MPEG-4 Book*. Prentice Hall, 2002.
6. T. Fourès and P. Joly. A multi-level model for 2d human motion analysis and description. In Internet Imaging IV, Simone Santini, and Raymondo Schettini, editors, *Proc. SPIE-IS&T Electronic Imaging*, volume 5018, pages 61–71, Santa Clara (CA), January 2003.
7. Y. Guo, G. XU, and S. Tsuji. Understanding human motions patterns. In *Proc. of International Conference on Pattern Recognition*, pages 325–329, 1994.
8. P. Tzouveli, G. Andreou, G. Tsechpenakis, Y. Avrithis, and S. Kollias. Intelligent visual descriptor extraction from video sequences. In *Proc. of 1st International Workshop on Adaptive Information Retrieval (AMR 2003)*, LNCS series. Springer, 2003.
9. C.R. Wren, A. Azarbayejani, T. Danell, and A.P. Pentland. Pfinder: real-time tracking of the human body. *IEEE Trans. Pattern Anal. Mach. Intell.*, 19(7):780–785, 1997.