

A multi-level model for 2D human motion analysis and description *

Thomas Foures and Philippe Joly

Institut de Recherche en Informatique de Toulouse
Université Paul Sabatier
118 Route de Narbonne
31062 Toulouse Cedex 4
France

ABSTRACT

This paper deals with the proposition of a model for human motion analysis in a video. Its main characteristic is to adapt itself automatically to the current resolution, the actual quality of the picture, or the level of precision required by a given application, due to its possible decomposition into several hierarchical levels. The model is region-based to address some analysis processing needs. The top level of the model is only defined with 5 ribbons, which can be cut into sub-ribbons regarding to a given (or an expected) level of details. Matching process between model and current picture consists in the comparison of extracted subject shape with a graphical rendering of the model built on the base of some computed parameters. The comparison is processed by using a chamfer matching algorithm. In our developments, we intend to realize a platform of interaction between a dancer and tools synthetizing abstract motion pictures and music in the conditions of a real-time dialogue between a human and a computer. In consequence, we use this model in a perspective of motion description instead of motion recognition: no a priori gestures are supposed to be recognized as far as no a priori application is specially targeted. The resulting description will be made following a Description Scheme compliant with the movement notation called "Labanotation".¹

Keywords: Motion analysis, human body model, motion description

1. INTRODUCTION

Automatic analysis of human motion in video sequences is a domain in wide expansion over the past few years. The reasons for such a development are particularly due to the high potential of exploitation through applications such as surveillance systems, user interface, or, in cultural domains such as sport or dance motion analysis. One of the most critical part in the conception of a system providing this kind of service is the way the human body is modelled, this decision being essential for the implementation. Such a model must be compliant with several requirements imposed by the application and its context. In our case, we will have to deal in real time with a video flow of low quality. Therefore, we propose a model which can be adapted itself to real time and noisy source conditions, and used to produced motion descriptions at different levels of details.

Many different human models already exist, such as H-Anim from the MPEG-4 standard. Meanwhile, this kind of model has been designed for synthetic purposes and so, does not present a total adequacy for its exploitation in an automatic analysis tool. For exemple, the few means of expression of physical constraints with H-Anim increases the difficulty of an automatic recognition task. Among the other kinds of model that have been proposed, statistic ones do not provide results with a high degree of precision, and so are not appropriate for an application where accurate motion descriptions are required.

In this paper we propose a multi-level modelling for human body. The principle of this modelisation is to deal with only one model for any kind of video document: it can adapt itself to the current resolution (or an expected level), and in that way reduce the computation cost and prevent for some wrong matching (as it may happen when model is defined with more details that it's possible to extract from frames). This model is

* in Proc. IS&T/SPIE Electronic Imaging 2003, Internet Imaging IV, Vol. 5018, pp. 61-71, Santa Clara (CA), 21-22 January 2003.

based on regions, described with ribbons. These ribbons can be decomposed in “sub-ribbons” in order to adapt themselves to the needs, implying a descent into hierarchical levels. Matching process between model and frames extracted from video flow requires at the first step a decomposition of studied pictures in research fields for each component of the model. Then a distance transform is applied on those areas in order to obtain a distance map. The comparison can be processed by using a chamfer matching algorithm, allowing to reduce matching error for each component.

The hierarchy property of the model has two majors consequences. The first one is its capacity to deal with real time processing. Indeed, a result, even if it’s at a coarse degree, can always be produced. The accuracy level depends on the time allowed to realize the processing, a short one leading to a low precision, a longer one to a higher. The second consequence is the system capacity to adapt itself to application needs. A user can specify the detail level according to his goal (if the description can be little developed, or on the contrary, requires a high precision), and then the appropriate model decomposition is used in consequence (reducing in the same way useless computation cost).

2. SHORT STATE OF THE ART

Most of the time, modelisation for human motion analysis consists in synthetic or statistic shape models. We can also find some other tools to describe motion based on optical flow,² pixel groups following,³ or point distribution model.⁴ Whatever, these tools are limited to more restrictive domains than global human motion analysis (such hand gesture recognition for exemple). Their usage for human motion description in real time seems to be difficult because of related imposed restrictions (in terms of framing, background, camera motion) and technical means that are required (optical capture tools, 3D set modelling, ...).

Those restrictions justify the development of the human body modelisation approach. The most classical ones are biomechanical models such as the one developed in MPEG-4 mentioned before. It is build with 2 kinds of objects: “sticks” which are linked by “joints”. This modelisation, based on the idea that articulations are centers of rotation, presents some limitations. It has been developed for synthetic purposes in order to animate avatars. Following this idea, motion possibilities should be as large as possible. A lot of detailed motions or positions which can be described with this model can not be take into account by motion analysis tools. Some of them can not be performed by an actual human body. Further more, no constraints can be specified on rotation angles, velocity, centers of gravity, Those constraints are important for analysing tools in order to filter impossible motions or to solve some ambiguous cases. Nevertheless, this approach has been followed in some works, in particular by Guo et. al.⁵ who are matching the skeleton obtained from the silhouette of the studied subject and their model, composed of segments linked with each other. We can easily imagine the complexity of such a task in a general case without any external and complementary knowledge. From the statistical point of view, we can find the works developed by Wren et. al.⁶ who used a model built with blobs (one per segment) in their system called PFinder. The background/object segmentation, and different body parts detection are performed on the base of some properties of spatial and color distributions of pixels in the picture. Even if the resulting description is actually conform with the original picture, its ability to take into account some details is strongly limited and the results quality depends on conditions in which the video has been recorded.

We propose to develop a model in order to overcome limitations identified for these examples. This means that the model must be generic, able to adapt itself to any kind of document, with a total independance with the environment in which the video has been shot, and easy to implement.

3. HIERARCHICAL MODEL

3.1. Main Orientations

We assume that a tool to analyse human motion, providing simple API’s for interaction (with an end-user), for explicit description requires the modelisation of a human body in a graphical way. As we mentioned before, models developed for synthetic purposes or statistic ones present some limitations. The solution we propose is hierarchically defined. The different levels correspond to levels of details we want to obtain as a result of analysis, or which best feat the video resolution. A first processing is performed on the first level of the model providing rough results. Then, these results can be refined by the application of a second level, where the description of

each body segment is sharper. On segments where results are better with this second description, we can apply a new level of refining, and so on until the results after the application of a new level are not significantly better than before. The final results will be the one which feat the best the resolution of the video.

We can distinguish 3 types of possible applications. First of all, the case where no knowledge on the studied video is available unknown and no real-time processing is required. Then a descent through all hierarchical levels can be realized, where only best matching is retained. This is the basic application. A second approach, incorporating more constraint, is also conceivable. As it has been mentioned before, the hierarchical implementation is, in its conception, highly compliant with real time processing. Indeed, time limitations can restrict the accuracy of the produced description but may allow the fact that a result (maybe quite coarse in comparison with the detail level which can be extracted from the video) can be produced. This kind of property can not be offered by a non evolutive model where a knowledge on all the components coordinates is required to define a subject position. Matching on streamed video in the general case becomes possible. The third allowed kind of use is the possibility for the user or the application to specify the accuracy of the desired level. In the case of a video in which, for example, a sharp description of arms movements and, in the same time, a coarser description of the legs are required, the model can be composed of high level components (in the hierarchy) for the top of the body and low level ones for the rest of the limbs. In that way, an adaptation to the actual needs can be achieved.

3.2. Model Definition

On the base of those general orientations, we have to define such a model. There are several possibilities to graphically design it. The different features which can be used are: sticks and joints association, edges, regions, or even blobs (for some statistical methods). The choice of these features is essential for the forthcoming process which will depend on it.

In our case, we propose to build a region-based model. The surface covered by one region of the model is supposed to match the surface covered by a body segment. The model is actually composed of a set of ribbons, which are associated with a given body segment. Those ribbons can be decomposed into "sub-ribbons" in order to provide the required level of details. Parameters used to describe those ribbons are not only their length and their width. The coordinates of a control point are also available in order to localize the segment in the picture. Those control points are generally located on corners of ribbons. An angle value (in fact the angle between a segment and the vertical axe) is added to this feature. Thus, 2 parameters are sufficient to locate a segment in an image (position and angle), plus 2 to generate it (length and width). Increasing the resolution (in terms of number of components) will inevitably raise the number of required parameters. This highlight the necessity to work with a model adapted to the frames resolution or to the accuracy level whished by a user, by limiting its parameters number to what is strictly necessary.

We propose at the moment 3 levels of description for the model (see Fig. 1). The first one corresponds to the coarsest definition of the studied subject. Here, only 5 ribbons are used: one for each limb, and one combining head and torso. On the second level, each segment splits up, on joints area, into two new ones. This new representation allows detection of limbs configurations that could not be evaluated at the previous level, such as bendings. Finally, a third level is proposed. It allows the distinction of the hands and feet by splitting up the model into components corresponding to the forearms and the down part of each leg. Obviously, this number of different levels is not exhaustive, and other developments can be made. Nevertheless, for our first experiments, we confine ourselves to this three representations, corresponding to a quite complete description of the human body.

4. MODEL/SUBJECT MATCHING

The purpose of this part is to describe extracted video features and methods used to employ the defined model. We propose to perform a model/subject matching in order to find for each component of the model the best location in the studied picture corresponding to the subject limb. The obtained positions for those components indicate the subject posture. We use the hierarchical decomposition of the model in order to perform, on a first step, the model/subject matching with the first model level. Thus, this operation produces some rough information about the limbs localization which can be employed during next steps when more refined levels of

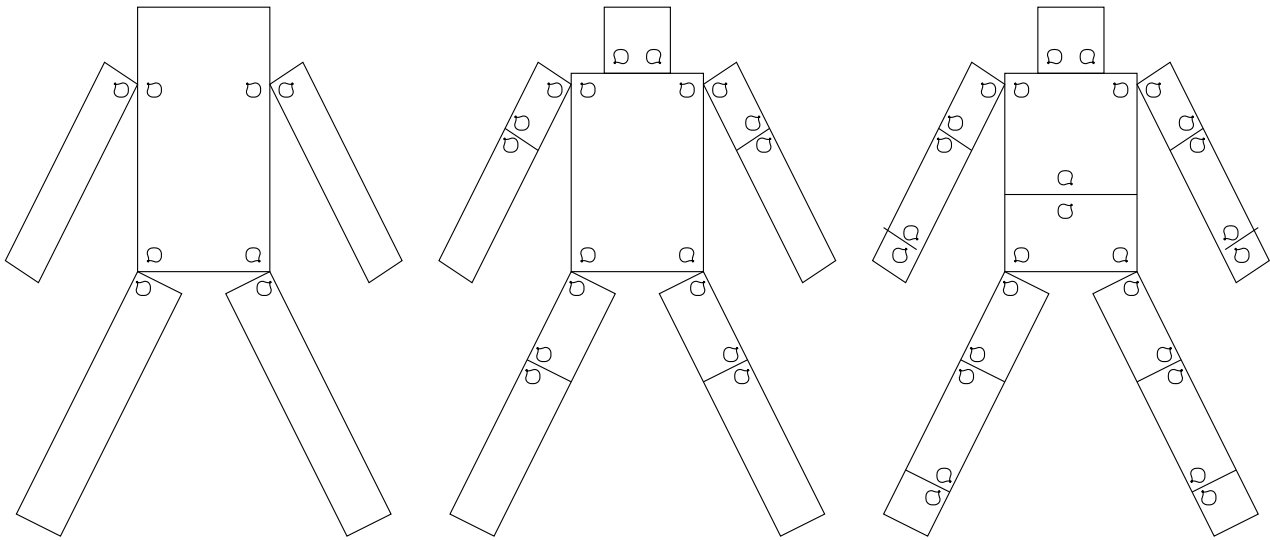


Figure 1. From left to right: the 3 model representations proposed.

the model are involved. This matching with a coarse model definition allows a reduction of the computational cost by reducing the research field for the more precise definitions of the model.

4.1. Silhouettes Pictures

The proposed model being region-based, the first step of preprocessing consists in features extraction from the current frame in accordance with this property. In our case, we have to obtain for each picture from the video, an image of the subject silhouette. To achieve this, the difference between a predefined background picture and the current frame is computed. Then noise is filtered using thresholds and morphological operators. The production of those silhouette pictures is followed by the creation of boundary boxes (see Fig. 2). The aim of this last operation is first to reduce the research field in an image containing only the region of interest (this second image having a smaller size than the first one), and also to define with the length/width ratio of the boundary box the size of the elements composing the model.

4.2. Model/subject Matching

4.2.1. Chamfer matching algorithm

The problem now is to determine from silhouette pictures obtained on the previous step, the determination of the correspondance between model components and subject limbs. In order to proceed to this, we propose to use the chamfer matching algorithm which principles have been exposed by Barrow:⁷ find the best match between two images features, this features could be edges, corners, or significant points. The method consists in the construction of a distance map computed with the chamfer distances on the image where the searched feature is located. 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. converting a binary image to a distance image) from the silhouette picture is processed by using the 3-4 Distance Transform defined by Borgefors,⁸ allowing a good approximation for the chamfer in only two passes over the image. Then, distance between the obtained map and the image of the searched feature is computed by using this one as a surperposed mask over the first. Several average measures for the value of the pixels located under the mask are possible, as arithmetic average, median value, or even maximum value. Meanwhile, we decided to use the root mean square (r.m.s.), presented by Borgefors as the one providing the less false minima.



Figure 2. Examples of 3 silhouettes obtained from a video sequence. Model/subject matching is realized from this kind pictures.

4.2.2. Implementation

We propose to implement the search of all model components positions in a sequential way: several successive coordinates are viewed and only the one providing the minimum r.m.s. value is selected.

In our case, the feature we are looking for in the silhouette image is region-based, which slightly differs works using chamfer matching algorithm in order to confront edges. In most of these cases, feature orientation is not a problem: it is a priori known and only information about localization within the image is requested. Here, moreover, we will also have to determine the subject limbs orientation (which is absolutely unknown). Thus, 2 parameters will be the points of interest: segments spatial coordinates and angles between segments one of the two axis. In that way, the search for the position minimizing the r.m.s. value has to take account into all the different possible orientations for one given model component, and for each of those orientations, find the best position in the image. The implemented algorithm is described by Fig. 3. From a given initial orientation, matching using only translations is processed (i.e. with a same angle value). Then a rotation of $\pm \frac{\pi}{4}$ around the center of the intersection area between the component image and the subject silhouette is applied. A new matching considering translations only for those angle values is performed. At this step, the best match for the three possible orientations is determined (initial angle and initial angle $\pm \frac{\pi}{4}$). We keep among these three possible cases the one associated with the more little r.m.s. value. Then the described process is reiterated, this time using a rotation value for the angle equal to the precedent divided by 2. A comparison between the previously computed position and the two determined at this step is evaluated, and only the best one is kept. The operation is repeated until the angle of rotation is updated with $\pm \frac{\pi}{128}$, which corresponds to 6 iterations and allows a low positioning error at the pixel scale in a regular video format picture.

The choice at the first iteration of a value equal to $\pm \frac{\pi}{4}$ for the rotation angle is due to the fact that elements composing the model are symmetric and then allow us to limit the research field by dividing it by two, reducing in the same time the computational cost. The choice of this initial value and the implemented algorithm ensure us to study every possible orientations for each model component. Moreover, the algorithm convergence is based on the fact that matching error will be minimal when an extracted segment from the model is globally oriented in the same direction than the corresponding segment located in the original image. Matching using only translations ensures limitation of effects due to potential neighbourhood of other segments.

As mentioned by Borgefors, chamfer matching algorithm has the disadvantage of leading to a potential false detection corresponding to a local minimum, this situation being likely to happen when the initial position

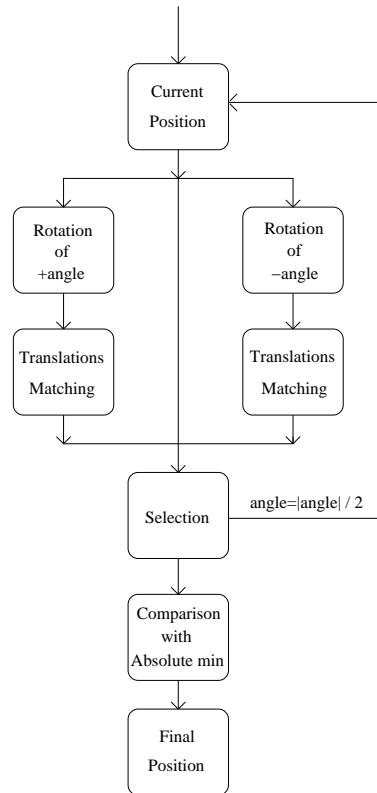


Figure 3. The different steps of a model component positioning in an image.

of a segment is too far from the optimal one in the image. To avoid this kind of situation, we propose to define for each component of the model a research area. This decision has two consequences. The first one is obviously to avoid a local minima, while the second one is, by reducing the research field within each image, to decrease processing time. Thus, the system becomes more efficient in terms of computational cost and in terms of reliability. In counterpart, there is the possibility to not have the researched limb located in its affected area, detection becoming in that case impossible. Regarding to the zones definition, and considering that no a priori information about subject posture is available, they are built according to the probable limbs localization in the image. The areas used in our implementation are given by Fig. 4.

5. EXPERIMENTAL RESULTS

Here are some matching results obtained in the case of the first level of decomposition of the model (see Fig. 5 and 6). It is important to precise that the matching process has been realized considering all model components independently one from the others, without applying any physical constraint on links or angles between the segments. Moreover, no information concerning matching on a precedent frame has been used, the pictures being also processed independently. Thus, research areas do not evolve during time, are not redefined by taking into account motion directions.

Considering the highest level of the model, we can see that matching results are, in an overall view, quite near to the real subject posture. Small shifts appearing at certain places are due to the presence in the research areas of pixels that do not belong to the concerned limb or, to the presence of holes in the subject silhouette, increasing non desired pixels weight. A common case where a matching can be considered as wrong is when limbs configuration can not be analysed by the model level. For example, a folded up leg as in the third picture enters this case. Of course, 2D human modelisation and pictures coming from only one camera present some

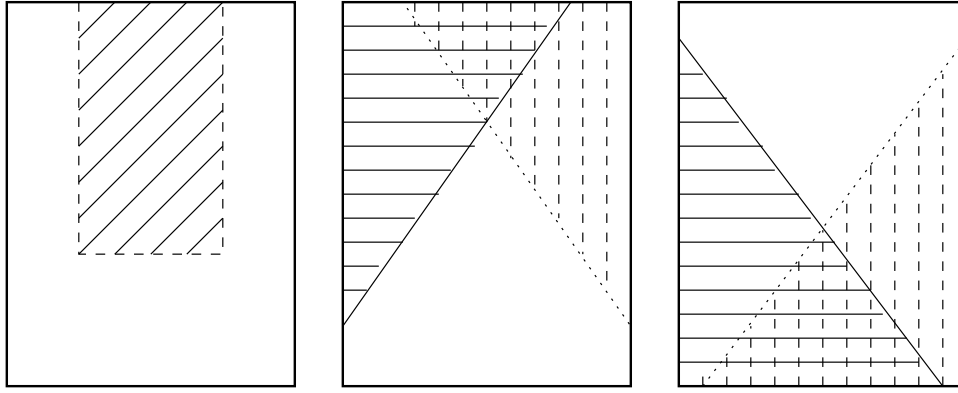


Figure 4. Image cutting in research areas. From left to right: torso research area (dotted lines), arms research area (right arm: full lines, left arm: dotted lines), legs research field (right leg: full lines, left leg: dotted lines).

limitations on some given subject postures analysis (in particular the ones where depth of field is important or where the subject is in profile) because 3D information is missing. Those cases will need data concerning the motion in its globality and the ability to follow limbs pixels to allow a good posture recovering.

A quality measure for the matching is proposed. Its objective is to provide an indication on the matching quality in order to take it into account or to reject this information. Used formula is given below:

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

This Q coefficient is a real value between 1 and -1. -1 means that the information is wrong, 1 means that the information is highly reliable. It will be used in further developments in order to take a decision on taking into account (or not) the current matching for the posture determination in the next pictures. We have to mention here that it is unlikely to obtain a value equal to 1 for Q because this would mean that all the points of the model cover pixels belonging to the subject. However, the model construction make this kind of situation impossible because all of the different parts are defined with a slightly bigger size than the one we may expect for the subject in order to insure some covering properties. A remark concerning Q is that this mesure provides some knowledge of matching quality in terms of surface covered, and does not indicate wether the limb detected is indeed the searched one or not. Only constraints installation in further works, in order to control distance between different model components, will be able to ensure a certain posture validity. However, those constraints have to be used carefully and can not force a segment positioning towards an another place, even if this one seems more appropriate. Constraints will have to play a role in modelisation of tensions involved in the human body.

Table 1. Q coefficient values obtained for the six silhouettes on Fig. 6

	T	RA	LA	RL	LL
Sil.1	0.4887	-0.3603	0.6777	-0.2584	0.3384
Sil.2	0.5255	0.0240	0.2678	0.0388	0.3682
Sil.3	0.5719	0.0135	0.1329	0.2409	0.3117
Sil.4	0.1673	0.4058	-0.2433	0.1284	-0.3517
Sil.5	0.4957	-0.3648	0.3316	-0.3107	-0.0275
Sil.6	0.3633	-0.0049	0.0323	-0.0867	-0.2624

6. ON GOING WORKS

On going works are dealing mainly with the exploitation the hierarchical levels. This operation has to be preced by a redefinition of the search areas. Indeed, as we said earlier, no information about the obtained matching



(1)



(2)



(3)



(4)

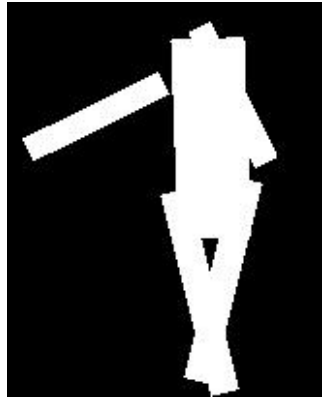


(5)

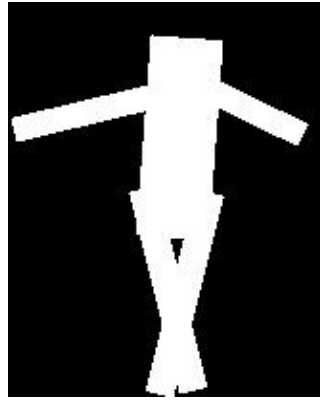


(6)

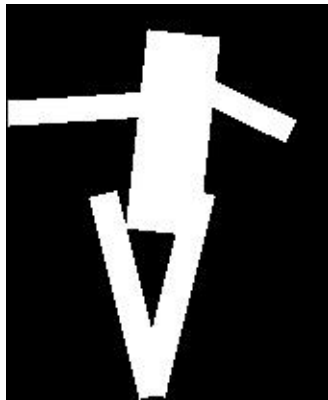
Figure 5. Frames extracted from a video sequence.



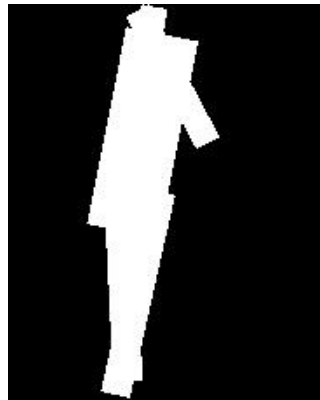
(1)



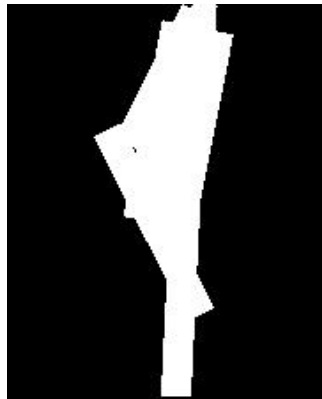
(2)



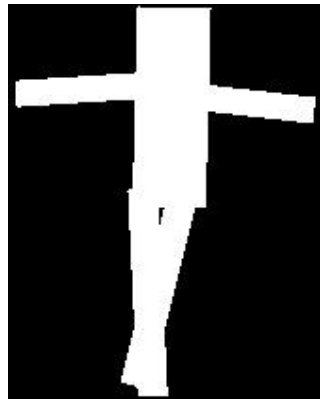
(3)



(4)



(5)



(6)

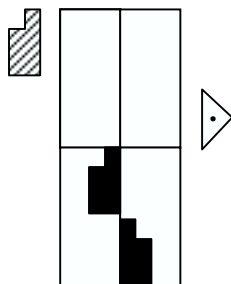
Figure 6. Matching results for frames given by Fig. 5.

on preceding frames in a sequence has been used in our algorithm. We employ the same definition for all the pictures and a taking into account preceding results during the current search should increase processing effectiveness, in terms of time (research field reduction) and reliability. We also have to add physical constraints on the space distribution of the different components of the model. These constraints should (with the help of some other heuristics) allow to pay more attention to positions meeting the conditions of distance while reducing the influence other ones. This should provide a dynamic motion modelisation and lead the search areas redefinition by giving a probable localization of the limbs in the image, and providing in the same time, some information on the possible validity of a posture.

7. DEVELOPMENTS

One of the main characteristics of the proposed model is to deal with real time processing. Furthermore, this work will be one of the elements of an interaction platform between dancers and tools of abstract movements conception. That is why we plan to use the model also for some description purposes. This motion description will be based on "Labanotation". This graphical formalism comes from dance, but has got the particularity to be very generic and so can be used to describe any type of human motion. By this way, each movement of each limb has a particular code, and no temporal segmentation is introduced at this point, relaying this task to a later specific process (using information concerning gesture recognition).

We propose to combine this motion notation with an XML-like description language. "A Labanotation-ML" language can be easily developed, using for example each unitary movement of each limb as an element with some attributes for its starting and ending position for example. In that case, motions can be considered as a group or a sequence of described limbs movements.



```

<RightFoot Id="rf1">
  <Motion Id="rf1m1" start_time=0 end_time=0.5 vertical="down" horizontal="front"/>
</RightFoot>
<LeftFoot Id="lf1">
  <Motion Id="lf1m1" start_time=0.5 end_time=1 vertical="down" horizontal="front"/>
</LeftFoot>
<RightArm Id="ra1">
  <Motion Id="ra1m1" start_time=1 end_time=1.4 vertical="center" horizontal="right"/>
</RightArm>
<LeftArm Id="la1">
  <Motion Id="la1m1" start_time=1.5 end_time=2 vertical="top" horizontal="front"/>
</LeftArm>

```

Figure 7. Possible XML transcription of a labanotation example

8. CONCLUSION

This paper proposed a multi-level model developed for a 2D human motion analysis in a video. The main motivation for this kind of modelisation is to produce valid information whatever the studied video document.

The main property of this model is to adapt itself to the image resolution, or to the application or the end-user needs, without any external intervention. There are four major consequences: this model definition being very generic, it can be applied without any a priori knowledge about the video; the hierarchical structure is an important factor in reducing the processing load; a result can be provided at any step of the matching algorithm, even at a coarse definition, allowing real time processing; and finally, this model can adapt itself to end-user or -application needs in terms of precision.

The implementation has been realized by using a chamfer matching algorithm, only on the first model level for now, and by processing each component independently one from the others, without usage of any temporal information. Results at this first abstraction degree show the method possibilities and the potentialities of future developments. On going works are related to the implementation of sub-levels of the hierarchical model. This implementation requires to evaluate precedent matchings, to update search fields (in order to reduce processing time and to increase reliability), for each frame.

Future developments will be based on the obtained motion description. Thus, at this moment, none segmentation process designed to proceed to a motion recognition is required by the algorithm. It will be possible to introduce it at the last moment, according to an external knowledge for example. To achieve this, we will propose a language combining the labanotation with a language such as XML, which should provide some interesting properties such as generic motion description.

REFERENCES

1. R. Laban, *Laban's Principles of Dance and Movement Notation*, MacDonald and Evans Ltd, London, 1975.
2. R. Polana and R. Nelson, "Low level recognition of human motion," in *Proc. of IEEE Workshop on Motion of Non-Rigid and Articulated Objects*, pp. 77–82, (Austin), 1994.
3. B. Heisele, U. Kressel, and W. Ritter, "Tracking non-rigid moving objects based on cluster flow," in *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, pp. 257–260, (San Juan), 1997.
4. T. Cootes, C. Taylor, D. Cooper, and J. Graham, "Training models shape from sets of examples," in *British Machine Vision Conference*, pp. 9–18, September 1992.
5. Y. Guo, G. Xu, and S. Tsuji, "Understanding human motions patterns," in *Proc. of International Conference on Pattern Recognition*, pp. 325–329, 1994.
6. C. Wren, A. Azarbayejani, T. Danell, and A. Pentland, "PFinder: real-time tracking of the human body," in *Trans. Pattern Anal. Mach. Intell.*, **19(7)**, pp. 780–785, IEEE, 1997.
7. H. Barrow, J. Tenenbaum, R. Bolles, and H. Wolf, "Parametric correspondence and chamfer matching: two new techniques for image matching," in *Proc. 5th Int. Joint Conf. Artificial Intelligence*, pp. 659–663, Cambridge, MA, 1977.
8. G. Borgefors, "Hierarchical chamfer matching: a parametric edge matching algorithm," in *Trans. Pattern Anal. Mach. Intell.*, **10(6)**, pp. 849–865, IEEE, 1988.