Human Body Tracking by Monocular Vision

F.Lerasle, G.Rives, M.Dhome, A.Yassine

Université Blaise-Pascal de Clermont-Ferrand,
LAboratoire des Sciences et Matériaux pour l'Electronique, et d'Automatique
URA 1793 of the CNRS, 63177 Aubière Cedex, France

Abstract. This article describes a tracking method of 3D articulated complex objects (for example, the human body), from a monocular sequence of perspective images. These objects and their associated articulations must be modelled. The principle of the method is based on the interpretation of image features as the 3D perspective projections points of the object model and an iterative Levenberg-Marquardt process to compute the model pose in accordance with the analysed image.
This attitude is filtered (Kalman filter) to predict the model pose relative to the following image of the sequence. The image features are extracted locally according to the computed prediction.
Tracking experiments, illustrated in this article by a cycling sequence, have been conducted to prove the validity of the approach.

Key-words : monocular vision, articulated polyhedric model, matchings, localization, tracking.

1 Introduction

In the last ten years, many researchers have tried to locate human body limbs by video or cine-film techniques. Their methods can be classified in function of the type of markers which are used. We can mention the marker free and marker based methods.

With regard to marker based methods, digitized data analysis can be done from anatomical landmarks by the placement of markers on the specific joints. The body can then be modelled using a multi link chain model (Winter [1]). After matchings between the primitives of the model and the digitized data, joint centers are reconstructed (Yeadon [2]). Automatic tracking system (Elite 1989) are also widely used. This system combines the real-time analysis of video-images, the signals acquisition relative to the muscular activity and external forces. The markers are passive (reflector markers) or active (infrared blankers, electro-luminescent Light Emitting Diodes). The active methods make the recognition and the tracking of the markers easier because each LED emission can be analysed separately. Yet, the physical constraints are greater.

Using these kind of markers are problematic. The non-rigidity wrapping during movement causes a relative body/markers displacement and induces uncertainty in the results. Moreover, wearing this kind of marker is quite easy for the ankles and wrists, but is difficult for complex articulations like shoulders, knees, hips. Moreover, adding passive or active markers induces some psychological effects on the subject such as rigidity in movements. To obtain accurate measurement, it is better to reduce the constraints on the subject as much as possible.

With regard to marker free methods, a well known technique in image processing is based on model matching using Distance Transformation (DT). It is described for instance in [3]. The method consists in making a DT image in which the value of each pixel is the distance to the nearest point in the object. The optimal position and orientation of the model can be found by minimizing some criteria function of these pixel values. Persson [4] uses this method with a simple 2D model of a leg prothesis.

Like Persson, Geurtz [5] developed a method based to a 2D representation of the body. The body limbs of the model are restricted to elliptical curves describing the segment contours. The image feature, used in the movement estimation, is only constituted by the contour data. His method is interesting but sensitive to noise. Consequently, results obtained on real images are somewhat inaccurate.

To solve the ambiguity (movement-depth), some researchers have proposed volumic models based on a priori knowledge of the human body. Rohr [6] introduces a model based approach for the recognition of pedestrians. He represents the human body by a 3D model consisting of cylinders, whereas for modelling the movement of walking he uses data from medical motion studies. The estimation of model parameters in consecutive images is done by applying a kalman filter. Wang [7] gives models of the different body limbs with simple geometrical primitives (such as cylinders, planar surfaces...) connected together by links which simulate articulations. The different images of the sequence are divided into regions by motion segmentation. From these detected regions and an affine model of the a priori movement, Wang deduces the 2D movement in each image of the set. With the knowledge of the volumic model, he estimates the 3D parameters of the model pose.

The modelisation error, due to representation with simple geometrical primitives, can have a negative effect on the interpretation result. A more precise modelisation should improve the results of the analysis.

In Robotic field, very few papers address the current problem which corresponds to the estimation of the spatial attitude of an articulated object from a single perspective image. Mulligan [8] presents a technique to locate the boom, the stick and the bucket of an excavator.

Kakadiaris [9] presents a novel approach to segmentation shape and motion estimation of articulated objects. Initially, he assumes the object consists of a single part. As the object attains new postures, he decides based on certain criteria if and when to replace the initial model with two new models. This approach is applied iteratively until all the object's moving parts are identified. Yet, two observed object's moving parts are supposed to be linked by only one inner degree of freedom.

2 Aim of the method

The research deals with the automatic analysis of 3D human movement based on a vision system. Like most of the developed methods, we need some prerequisites which are the knowledges of the observed object, free matchings between 2D image primitives and 3D model elements, and assumption about the projection of the real world on the image (perspective in our case). Our analysis will be based on the articulated volumic model and a localization process which computes the attitude of the 3D object model such that the selected model elements are projected on the matched 2D image primitives. A similar approach can be found in [10] where Lowe proposes a technic to fit a parametrized 3D model to a perspective image. Nevertheless, our method differs from Lowe 's one in the minimized criterion: Lowe uses a 2D criterion calculated in the image plane. Our criterion is 3D one which permits to greatly simplify the computations involved.

To give more precision about the model, the human body model will be deduced from Resonnance Magnetic Imaging (RMI) measurements and will be manipulate like a set of rigid parts which are articulated to each other. A cycling sequence was taken to illustrate this research.

3 Method

3.1 Model description

The modelisation of the human body is similar to the results of work carried out in the laboratory on articulated objects (Yassine [11]). We describe them briefly.

In our system, the articulated object comprises several Computer Assisted Design models (CAD) connected by articulations which describe the possible relative displacements between parts. Each CAD model corresponds to the polyhedrical approximation of the real part of the object. It is composed by a set of 3D vertices, a set of links between vertices to define the ridges and a set of links between ridges to build the different faces.

Each articulation is characterized by a set of degrees of freedom. Each inner degree of freedom is defined by its type (rotation and translation) and its 3D axis.

To animate the model, we have defined operators which allow us to place, in the observer frame, a 3D model vertice which initially defined in the model frame (see Yassine [11]).

3.2 Matchings 2D-3D

Before the localization process, we must extract some image features and match them with the associated geometrical primitives of the articulated model. Dhome [12] computes the pose of a simple polyhedrical object from matchings between all the visible model ridges and segments extracted from grey images. In our application, where the CAD model associated with each part (shank, thigh...) has a repetitive structure (comparable to a skew netted surface), such an approach is not applied. Only ridges, which are limbs, will be matched. A limb is a ridge common both to a visible surface and an invisible surface, after projection of the model in the image plane. Obviously, the non-rigidity of the bodily wrapper, during the movement, causes an incoherence comparatively to the static model (and so comparatively to the detected limbs). Yet, for smooth movement (like pedaling), deformations are quite insignificant.

To bring more constraints, some random points are chosen on the model surface and are matched with their features detected in the image. This tracking of specific points is comparable to a classic technique of marker tracking, but the advantages are obvious : the number and the location of these characteristic points are not pre-defined compared to markers. These points can be ignored or replaced by other random points during the tracking process.

The primitives extracted from the image are also straight segments or points.

Matchings 2D-3D from points: Given the model pose relative to the image I^1, some 2D points p_i^1 are selected in this first image and their 3D coordinates are deduced by reverse perspective projection. The equivalents p_i^n ($i = 1..p$) of the points p_i^1 are searched in the consecutive images (noted I^n) of the sequence by analysis of the images grey levels.

A priori, it is necessary to search the points p_i^n in a sufficiently large zone of the image I^n to include the displacement existing between the homologous points of the two images. To reduce the combinative, we take the predicted position estimated for the image I^n into account. This prediction will be computed by a Kalman filter ([13], [14]). The points p_i^n in the image I^n will be researched in proximity of the predict projected point (noted $p_i^n pred$) of the model. Each point p_i^n is obtained after maximizing correlation scores on grey levels windows included respectively in images 1 and n (Lerasle [15]). In fact, the predict pose makes it possible to restrict the research domain for the correlation and thus to reduce the processing cost.

To sum up the method, some 3D points of the model are associated to textured figures (centred at points p_i^1) of the initial grey level image. These points are localized in the following images of the sequence after correlation on the grey levels. The localization of these 2D points,

in the different images of the sequence, will be better if the texture included in the images is rich. For this reason, during the film, the observed subject is wearing a pair of tights with non repetitive texture.

Moreover, to manage the possible vanishing of these 3D points and thus the textured designs associated with the initial image, this list of 3D model points will be modified during tracking. From one image to another, from the computed attitude, some points are removed from the list and replaced by others chosen randomly in their proximities to respect the initial spatial spread. This modification is managed like a stack Last In First Out so that after n localizations corresponding to the n first images of the sequence, all the points in the initial list have been removed. It is in fact, a markers system (or rather textured designs markers) sliding.

Matchings 2D-3D from limbs: At step n of the tracking (image I^n), for the attitude according to the image I^{n-1}, a Z-buffer or depth-buffer method (Catmull [16]) allows to extract the model ridges which correspond with limbs.

The idea of wearing a textured dark pair of tights moving in front of a white background can also be used. Projections of the points extremities (noted $A_i^{n-1}B_i^{n-1}$) of the limbs L_i are removed on to the white/dark transitions of the image I^{n-1}. Then, the correspondents $A_i^nB_i^n$ (in the image I^n) of the points $A_i^{n-1}B_i^{n-1}$ will be deduced after correlation on the respective grey levels. The segment $A_i^nB_i^n$ will be matched with the limb L_i and so with the 3D associated ridge of the model.

3.3 Localization process

We describe the problem of the localization of an articulated object, give briefly the mathematical equations and the algorithm used to solve them. The validity of this algorithm was proved by Yassine [11] on articulated polyhedrics objects like an operator arm. The pose estimation of an articulated object from a monocular perspective image depends on $10 + q$ parameters. The first four ones are the intrinsic camera parameters. The six following ones are the extrinsic parameters (noted $\alpha, \beta, \gamma, u, v, w$). These parameters correspond to the three rotations and the three translations around and along the observer frame axis which permit to locate a rigid object. In the present process, these parameters determine the pose of the reference part of the viewed articulated object. The q following ones (noted $a_1, ..., a_q$) represent the inner degrees of freedom.

The problem we intend to address can be described as follows. We suppose known : the perspective projection as a model of image formation, the intrinsic parameters of the acquisition system, the CAD model of the viewed object and a sufficient set of matchings between image features and model primitives.

Then, we must obtain the $6 + q$ parameters which define the object location minimizing the sum of the distance between the matched model primitives and the interpretation planes(planes through the optical center of the camera and the considered segments). Thus, for the model primitives like ridges, only the extremities of these ridges are considered. For each matched ridge, we minimize the sum of the distance between the two extremities of the model ridges and the correspondent interpretation plane. For the matchings on points, the matched points in the image ($p1$ in figure 1) will be replaced by two segments which will be perpendicular in the image. Then, we will minimize the distance between the 3D point ($P1$ in the figure 1) of the model and these two perpendicular interpretation planes.

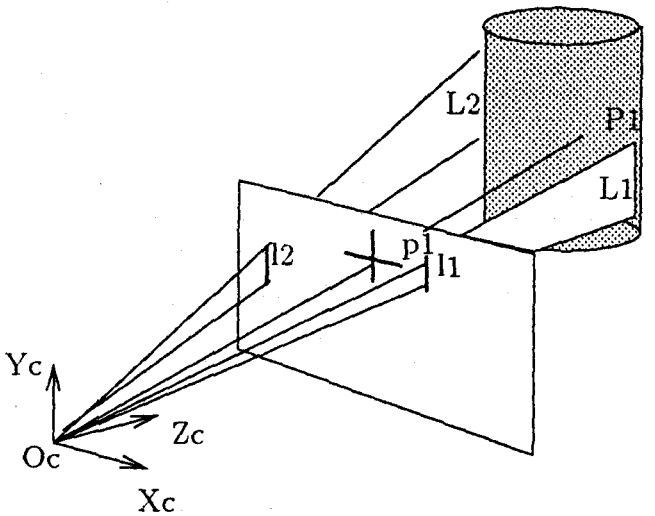

Fig. 1. example of interpretation planes

Distances to interpretation planes: In the following equations, the index c refers to the observer frame, m refers to the model frame and the variable A refers to the position vector with $A = (\alpha, \beta, \gamma, u, v, w, a_1, .., a_q)$. Let p image segments l_i be matched respectively with p model points P_i^m. It was assumed that all the vectors and points are expressed in the camera coordinate system. To compute the vector A, we merely express that the transformed point P_i^c by the transformation represented by A of the point P_i^m must lie in the interpretation plane Π_i (normal $\vec{N_i}$) of the corresponding image segment. This can be written by a product scalar function. A method to linearize this function consists in approximing this function near the value $(A)_k$, A at step k, by a first order Taylor development :

$$F(A, P_i^m) = (\vec{N_i}.\overrightarrow{OP_i^c}) \approx F(A_k, P_i^m) + \frac{\partial F(A_k, P_i^m)}{\partial A}(A - A_k)$$

where $i = 1..n$ is the index of the matched point (n the number of matched points).

Resolution of the system: If A is the solution, $F(A, P_i^m) = 0$ and thus the set of 3D matched points allows to construct a system with n linear equations : $[J]_k^T.(E_k) = [J]_k^T.[J]_k.(\Delta A)_k$ with

$$(\Delta A)_k = \begin{pmatrix} \alpha_k - \alpha \\ . \\ . \\ a_{1k} - a_1 \\ . \\ . \\ a_{qk} - a_q \end{pmatrix}, (E)_k = \begin{pmatrix} F(A_k, P_i^1) \\ F(A_k, P_i^2) \\ . \\ . \\ F(A_k, P_i^n) \end{pmatrix}, (J)_k = \begin{pmatrix} \frac{\partial F(A_k, P_1^m)}{\partial \alpha} & \cdots & \frac{\partial F(A_k, P_1^m)}{\partial a_q} \\ \frac{\partial F(A_k, P_2^m)}{\partial \alpha} & \cdots & \frac{\partial F(A_k, P_2^m)}{\partial a_q} \\ . & \cdots & . \\ . & \cdots & . \\ \frac{\partial F(A_k, P_n^m)}{\partial \alpha} & \cdots & \frac{\partial F(A_k, P_n^m)}{\partial a_q} \end{pmatrix}$$

This system will be solved by an iterative Levenberg-Marquardt approach [17]. By this way, the global criterion to minimize is :

$$Error = \sum_{i=1}^{p} (\vec{N_i}.\overrightarrow{OP_i^c})^2$$

At each iteration of the iterative process, we obtain a correction vector to apply to the position vector $(A)_k$: $(A)_{k+1} = (A)_k + (\Delta A)_k$

The iterative process is repeated until a stable attitude is reached meaning $Error < \epsilon$. In any case, the matrix of partial derivatives $(J)_k$ must be computed and these calculations will be set out in detail in [15]. The computation of the covariance associated with the model attitude have been detailed in [15] too.

4 Experiments and results on tracking

The matchings between 3D model primitives in the attitude linked to an image of the sequence and 2D primitives in the following image are automatic. At the step n of the tracking (image I^n), the matchings process takes the attitude computed for the image I^{n-1} into account. So, the attitude linked to the first image of the sequence and the set of markers linked to this image must be known in advance. In fact, initialization of the tracking process requires two steps :

1. manually, the operator superimposed the model to the first image of the sequence,
2. for this attitude, the operator manually selects some visible points from the model. A good spatial spread of the set of these points increases the constraints caused by these points.

For example, the following figure represents the first image of the cycling sequence and the markers which have been choosen.

Fig. 2. initial grey level image with selected 3D points projection

The tracking process have been validated on real images sequences. We have choosen to present the results obtained on a cycling sequence. First, we have built the shank and thigh CAD models of the cyclist. These models have been deduced from images provided by an RMI scan. This scan consisted in 34 cross-sections of the legs. Low level treatement (smoothing and contours detection) makes it possible to extract peripheral contours of each cross-section. The

contour points coordinates x and y of each cross-section, associated with its height z, enables us to determine the 3D vertices of the model (see figure 3).

Moreover, we had to compute the articulation model of the knee which is quite complex. An articulation model with three rotations, corresponding to the flexion-extension (axis Oy), to the internal rotation (axis Oz) and the valrus-valgus rotation (axis Ox) has been choosen.

The figure 4 represents the projection of the attitudes computed during the sequence from a point of view located in the cycling plan. The six following figures (end of article) represent the model projection superimposition on different images of the sequence. Obviously, the model surfaces which are not really compatible with the grey-level image target are distort surfaces like the back of the leg or calf.

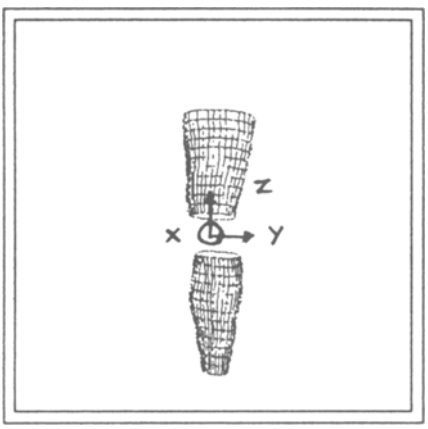

Fig. 3. leg model used for the pedaling sequence

Fig. 4. model tracking from a point of view situed in front of the bicycle

5 Conclusion

This paper presents a new method for estimating, in the viewed coordinate system, the spatial attitude of an articulated object from a single perspective view.

The approach is based :

- on the a priori knowledge of the visualised object meaning the knowledge of its model,
- on the interpretation of some image points as the perspective projection of 3D model points,
- on an iterative search of the model attitude consistent with these projections,
- on a calculation of the covariance matrix associated with these projections.

The presented method is quite different from the markers method because we don't use real but fictitious markers (through a textured pair of tights). Thus, the proposed method is more flexible because the number of these markers and their emplacements are not a priori fixed. Moreover, wearing a pair of tights causes no psychological effects, no physical constraints and no added techniques (comparatively to active markers method).

The inaccuracies of the method are in process initialization and especially in the approximate estimation of the model attitude compatible with the first image. This first pose has an effect on the quality of the localizations obtained during the tracking. Moreover, the defined articulated static model is not always consistent with the image target because the body wrapping is not always constant during the movement.

Our next purpose will be to improve both the initialization and the modelisation, and to analyze occulting movement, for example the occulting of one leg by one another. The implemented Kalman formalism should help us to cope with this kind of problem.

To our knowledge, such marker free method has never been applied to human movement and could be useful to further study biomechanical and energetics of muscular activity aspects.

References

1. A.D. Winter. A new definition of mechanical work done in human movement. *J. Aplli. Physiol.*, 46:79–83, 1979.
2. M.R. Yeadon. A method for obtaining three-dimensionnal data on ski jumping using pan and tilt camera. *International Journal of Sport Biomechanics*, 5:238–247, 1989.
3. G. Borgefors. On hierarchial edge matching in digital images using distance transformations. *Internal Report of the Royal Inst. of Technology, Stockholm*, 1986.
4. T. Persson and H. Lanshammar. Estimation of an object's position and orientation using model matching. In *Proc. of the sixth I.S.B Congress*, 1995.
5. A. Geurtz. *Model Based Shape Estimation.* PhD thesis, Ecole Polytechnique de Lausanne, 1993.
6. K. Rohr. Towards model based recognition of human movements in image sequences. *Image Understanding*, 59(1):94–115, January 1994.
7. J. Wang. *Analyse et Suivi de Mouvements 3D Articulés : Application à l'Etude du Mouvement Humain.* PhD thesis, IFSIC, Université Rennes I, 1992.
8. I.J. Mulligan, A.K. Mackworth, and P.D. Lawrence. A model based vision system for manipulator position sensing. In *Proc. of Workshop on Interpretation of 3D scenes, Austin, Texas*, pages 186–193, 1990.
9. I.Kakadiaris, D.Metaxas, and R.Bajcsy. Active part-decomposition, shape and motion estimation of articulated objects : a physics-based approach. In *Computer Vision and Pattern Recognition*, 1994.
10. D.G. Lowe. Fitting parameterized three-dimensional models to images. *PAMI*, 13(5):441–450, May 1991.
11. A. Yassine. *De la Localisation et du Suivi par Vision Monoculaire d'Objets Polyédriques Articulés Modélisés.* PhD thesis, Université Blaise Pascal de Clermont-Ferrand, 1995.
12. M. Dhome, M. Richetin, J.T. Lapresté, and G. Rives. Determination of the attitude of 3d objects from a single perspective image. *I.E.E.E Trans. on P.A.M.I*, 11(12):1265–1278, December 1989.

13. N. Daucher, M. Dhome, J.T. Lapresté, and G.Rives. Modelled object pose estimation and tracking by monocular vision. In *British Machine Vision Conference*, volume 1, pages 249–258, 1993.
14. N. Ayache. *Vision Stéréoscopique et Perception Multisensorielle.* Inter Editions, 1987.
15. F. Lerasle, G. Rives, M. Dhome, and A. Yassine. Suivi du corps humain par vision monoculaire. In *10th Conf. on Reconnaissance des Formes et Intelligence Artificielle*, January 1996.
16. E. Catmull. *A Subdivision Algorithm for Computer Display of Curved Surfaces.* PhD thesis, University of Utah, 1974.
17. D.W. Marquardt. *Journal of the Society for Industrial and Applied Mathematics*, 11:431–441, 1963.

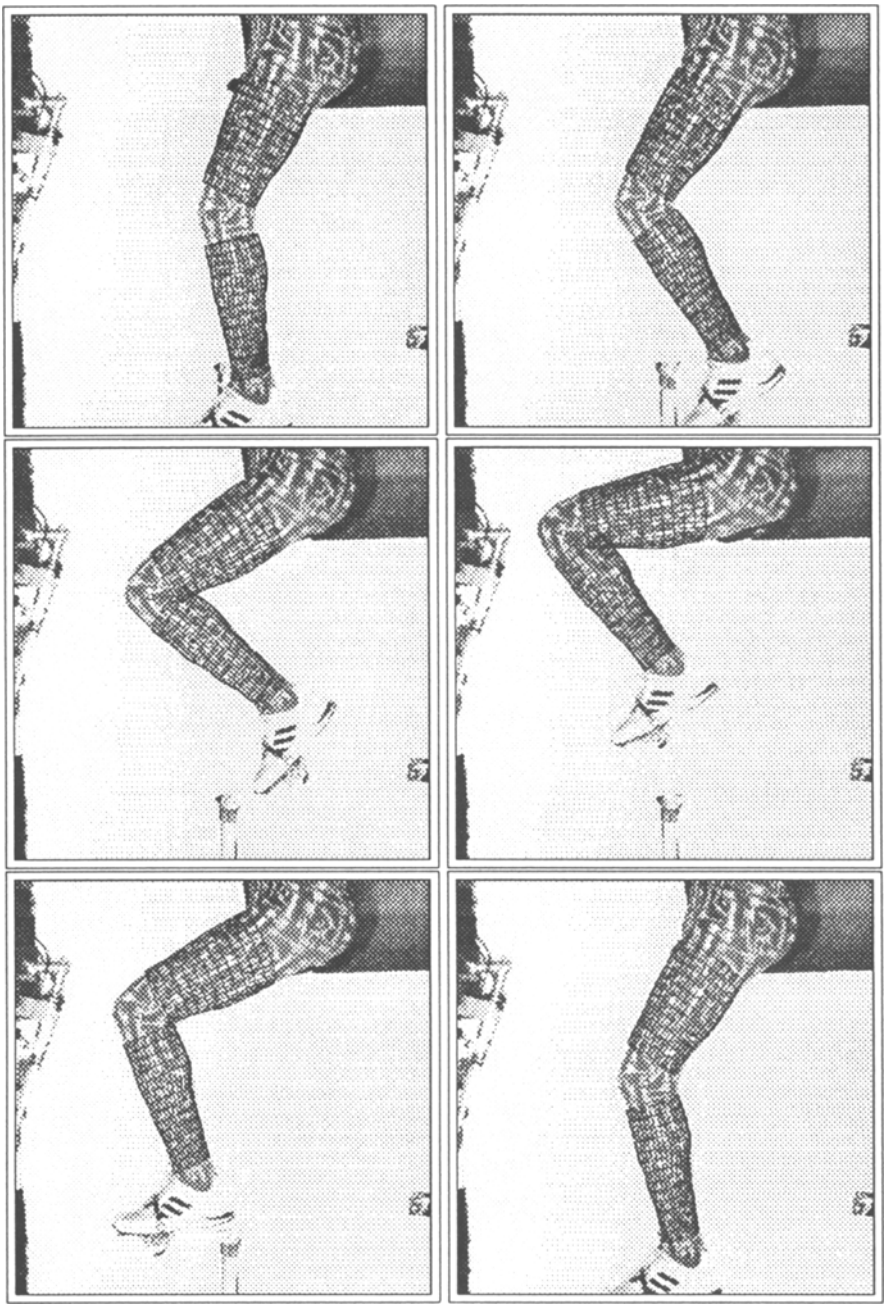

Fig. 5. model projection superimposed to the images 1,10,20,30,40,50 of the sequence