Motion estimation of transparent objects in the frequency domain

M. Pingault¹,*, D. Pellerin¹,²

¹Laboratoire des Images et des Signaux (Lis) INPG, 46 Av. Félix Viallet 38031 Grenoble Cedex, France
²ISTG, Université Joseph Fourier, Grenoble, France

Abstract

Motion transparency phenomena in image sequences are frequent but classical methods of motion estimation are unable to deal with them. So there is a need for more general techniques in order to solve this important problem. The method described here is based on an image sequence analysis in the frequency domain. It is mainly composed of a Stochastique-Expectation-Maximisation (SEM) algorithm which offers a new statistical model for this problem. This method, despite its large execution time, offers some interesting results on artificial and natural image sequences.

Key words: Velocity estimation, motion transparency, frequency domain, statistical model, SEM algorithm

* Corresponding author.
Email address: pingault@lis.inpg.fr (M. Pingault),
pellerin@lis.inpg.fr (D. Pellerin).

Preprint submitted to Elsevier Science 12 November 2002
1 Introduction

In a general case, motion estimation is useful to understand a scene. But classical methods, which are based on the assumption that there is a single motion in each part of an image, reach their limits when transparency phenomenon appear.

In nature, motions in transparency are very frequent. These phenomena happen when at least two elements are superposed in an image sequence and when these elements are animated by different motions. The most well-known examples of motion transparencies are reflections on transparent surfaces (e.g. windows, water). These phenomena can also be found in partial transparencies, aka partial occlusions. These kinds of transparencies happen when the object in the foreground is fragmented. The easiest examples are occlusions caused by gates or branches of trees. These two sorts of transparencies are considered as additive, that is to say the texture patterns of objects are sum to give the global texture.

Several methods have been proposed to trait the transparency phenomena. Shizawa and Mase [1,2] have formulated a model to resolve this with the help of the optical flow constraint equation (OFCE), often used in motion estimation method. They derive this equation to its higher order for each motion. Nevertheless, This method has only be tested on white noise image sequences. Those authors have, at the same time, exposed the bases of transparency estimation in the frequency domain. Then, in the frequency domain, two transparent moving objects are summary to two plans, called motion plan, whom parameters are the objects velocities.
From then, it is possible to abord the transparency phenomena by two way: by temporal methods (based on differences between successive images of a sequence) or by frequency methods (based on the amplitude or phase of the Fourier transform).

Among the temporal methods, Jepson and Black [3] have been the first who adopt a statistical approach to the transparency phenomena. The authors used the OFCE on a neighborhood of the pixel whom they looked for determining the velocity vector. This permits to obtain a set of lines depending of the two components of the velocity vector. This lines cut themselves into different points which correspond to the number of transparent objects. So, by a probabiliste mixture of gaussian, the extension of the Expectation-Maximisation (EM) algorithm, the authors are able to determine the value of the different velocities. Nevertheless, Their conclusion is a little reserved because the method are needed a great number of points and the constraints due to the noise are created, despite of the robust estimation some importante errors.

The number of frequency methods is much more vast. The nulling-technic of Darrell and Simoncelli [4] are used oriented filters which successively retains one of the transparent object (called layer) of the image sequence and, therefor, allows to estimate the remaining object velocity vector. The problem of this method is it needs a previous knowledge of one of the objects to be able to nulle (in the way of Darrell and Simoncelli) a layer and then the other in a recursive way. Another technic is based on some combinaisons, called triades, of Gabor filters [5]. This filtering gives an indetermination on the sign of the velocities, so limits its application to the image sequences in which the movie camera is not moving. Vernon [6] uses an equation system based on the Fourier transform of each image of a sequence. With the study of its phase, linear in
comparaison with the spatial frequencies, he deduces the velocities. The results shown are only particular cases because composing of image sequence moving with entire velocities. The method of Chen and al. [7] consists in consentante ones attention to the study of an particular axe of the cubic volume of the spatio-temporal frequency domain to consisiste à se concentrer sur l’étude d’un axe particulier du volume cubique des fréquences spatio-temporelles pour drill the different plans corresponding to moving objects. This manner reduce the calculus volume but it also leads to yield the method sensible to the lack of energy in some regions of motion plans.

The proposed method is efficient in the frequency domain. The Fourier transform permits to dissociate the transparent objects into motion plans. Their parameters (velocities) are estimated thanks to a statistical model around the Stochastique-Expectation-Maximisation (SEM) algorithm. First of all, we describe the theory of transparent phenomena in image sequence, focusing on the case of additive transparency. Then, we develop the used method to estimate the optical flow of an image sequence. Finally, we expose some results on artificial and natural image sequence.

2 Theory of transparent moving objects

The estimation of motion in the domain of frequencies is based on one fundamental result. This result becomes from analyze of a video sequence by a three-dimensional Fourier transform. Some hypotheses, describe along this section, are made to build this model.
2.1 Motion equation

To simplify the theory, we will first examine the case of one motion in each region of the image sequence, otherwise the case where each pixel of the sequence is animated by only one motion (without transparency). Of course, it can exist several motion in the whole sequence.

A video sequence can be shown like:

\[
i : \left( \mathbb{R}^3 \rightarrow \mathbb{R} \right),
\]

\[
(x, y, t) \mapsto i(x, y, t),
\]

where \(x\) and \(y\) are, respectively, horizontal and vertical image coordinates and \(t\) is time.

The first assumption is it exists in the sequence a function \(i_0(x, y)\) which is such as :

\[
i(x, y, t) = i_0(x - v_x t, y - v_y t).
\]

Due to this equation, moving objects must move with a uniform velocity vector \((v_x, v_y)\) and must have a constant illumination. Now, calculating the spatial and temporal Fourier transform of the sequence \(i(x, y, t)\), it provides :

\[
I(f_x, f_y, ft) = I_0(f_x, f_y) \delta(v_x f_x + v_y f_y + ft),
\]

where \(I_0\) is spatial Fourier transform of \(i_0\) and \(\delta\) is the Dirac delta function.

Thus, \(I(f_x, f_y, ft)\) is nonzero only in the plan, called motion plan, which contains frequency origin, its equation is:

\[
v_x f_x + v_y f_y + ft = 0.
\]
Estimate parameters of this plan leads to estimate velocity vector components \((v_x, v_y)\) of the moving object.

### 2.2 Transparency additive equation

We consider, now, it could exist multiple motions at a same location. That is to say in a pixel it could have several velocity vectors. Equations in the way of additive transparency become, simply, of equations in the upper case by the result:

\[
\begin{align*}
\text{If, } & i(x, y, t) = \sum_{k=0}^{p} i^k_0(x, y, t), \\
\text{with, } & i^k_0(x, y, t) = i^k_0(x - v^k_x t, y - v^k_y t), \\
\text{then, } & I(f_x, f_y, f_t) = \sum_{k=0}^{p} I^k_0(f_x, f_y) \delta(v^k_x f_x + v^k_y f_y + f_t).
\end{align*}
\]

This result is due to the linearity of the Fourier transform. The spectrum of such sequence is nonzero only on \(p\) motion plans passing by frequency origin.

\[
v^k_x f_x + v^k_y f_y + f_t = 0.
\]

Estimate velocities of transparent objects in the sequence is the same to estimate all components of these plans.

### 3 method

The principe of motion estimation method in the frequency domain is to calculate on each pixel of an image sequence the local spectrum, and then estimate
the parameters of each motion plan, which are velocities. This method can be divided in three steps: the filtering following by the calculus of Fourier transform, the extraction of local maxima in the spectrum and finally, the estimation of motion plans parameters.

3.1 Filtering and Fourier transform

Three choices have to be done before doing the Fourier transform.

Choice of the filtering. One can see with equation (6) that all motion plans pass through frequency origin so it is difficult to dissociate then in low frequencies. More, natural image sequences have, it is a mean, spectrums in $1/f$ in the three frequency directions. This decrease is thus rapid enough so the high frequencies have to be raise to fit correctly the plans. We choose, to do so, the inverse of the gaussian filter of size $3 \times 3$.

Choice of the window. The Fourier transform of the image sequence is calculated on a spatio-temporal window around the pixel to analyze. This window
has to be large enough so the Fourier transform is integrated an large enough motion. But, it also has to be as little as possible to respect the hypothesis of uniform translational motion. The compromise we experimentally choose is a window of size $15 \times 15 \times 15$.

**Choice of the padding.** This operation, made in order to over-sample the Fourier transform, is needed to improve the legibility of the spectrum but it increases the volume of calculus. We choose as compromise to carry out operations of passing in frequency domain with spatio-temporal data of size $32 \times 32 \times 32$.

### 3.2 Maximum extraction

The spectrum obtain in the previous section is not exploitable directly. Indeed, only points in the plans are wanted, but in spectrum of natural sequence, notably because of local segmentation, there are some frequencies which are out of plans and are nonzero (Fig. 2(a) and 2(b)). Then, to avoid those points which are outliers for the model we use an maximum extraction method. The current pixel is compared recursively with its eight neighborhoods along spatial frequency and with two neighborhoods along temporal frequency (Fig. 3) and with a threshold which deceases linearly up to have a number of points around 300 (Fig. 2(c)).

### 3.3 motion parameters estimation

We now dispose of a set of $N$ points $(f_x, f_y, f_t)$, subsequently called $X_{1\rightarrow N}$, distribute on cloud theoretically orientate following $p$ motion plans whose pa-
Fig. 2. (a) Example of two motion transparency segments ($v^1_x = -1$ et $v^2_x = 2$ pixels per image) in space ($t, x$). (b) Modulus of the real Fourier transform of the "one dimensional image" sequence in the space ($f_x, f_t$). (c) Points of the spectrum kept by the extraction of local maxima.

Fig. 3. Non-linear maximum filter
rameters are the unknown quantities. We thus model the frequency domain by a mixture of \( p \) plans by the way of defining \( \varepsilon_{1\rightarrow N} \) which represent estrangement errors of points \( X_{1\rightarrow N} \) to its corresponding motion plan. We suppose the density of probability of those errors is a gaussian one with zero mean and variance \( \sigma^2_\varepsilon \). This assumption can be made because the points are symmetrically distribute in both sides of motion plans.

The model we construct is a missing data one. We thus introduce latent data, called indicators variables of motion plan.

In order to solve the augmented model, we use the SEM algorithm which permit to maximize the density of probability a posteriori augmented. The formalization of this algorithm is based on the work of Cord and Declercq [8].

By now, we will note the indicator variables of motion plans \( \tilde{X}_{1\rightarrow N} \). For all \( j \) in \([1, N]\), \( \tilde{X}_j \) is a vector of size \( p \) which takes its value in the discrete space \( \{0, 1\}^p \) and which indicates at what plan the point \( X_j \) belongs. So the \( k^{\text{ième}} \) component of \( \tilde{X}_j \), noted \( \tilde{X}_j^k \), will be equal to 1, if the \( j^{\text{ième}} \) point belong to the \( k^{\text{ième}} \) plan. More, we suppose a point can only be in one and only one plan. Those vectors have thus only one component non-zero. Then, because of the error \( \varepsilon_j \) have a normal distribution, we can express the probability that a point \( j \) belongs to a motion plan \( k \) as following:

\[
\alpha_k = p(\tilde{X}_j^k = 1) = \frac{1}{\sqrt{2\pi\sigma^2_\varepsilon}} e^{\exp \left( -\frac{(v^k_x f_{x_j} + v^k_y f_{y_j} + f_{t_j})^2}{2\sigma^2_\varepsilon} \right)}. \tag{7}
\]

Usually, in mixture distribution problem, the evolution of those variables are defined by a multinomial distribution:
\[ \tilde{X}_j = \mathcal{M}(1; \alpha_1, \ldots, \alpha_p), \]  

where all \( \alpha_k \) are normalized.

Now by the knowledge of those indicator variables of motion plan, the last probability of density we have to explicit, it is the posterior one which is found by the Bayes theorem:

\[
\begin{align*}
\mathcal{L}(v^k_x, v^k_y, \sigma^2_x | X_{1-N}, \tilde{X}_{1-N}) &\propto p(X_{1-N} | v^k_x, v^k_y, \sigma^2_x, \tilde{X}_{1-N})\pi(v^k_x, v^k_y, \sigma^2_x).
\end{align*}
\]

The density of probability \( p(X_{1-N} | v^k_x, v^k_y, \sigma^2_x, \tilde{X}_{1-N}) \) is the conditional likelihood of the augmented model. The spatial independence of data allow to develop this likelihood as the product of terms \( p(X_j | v^k_x, v^k_y, \sigma^2_x, \tilde{X}_j) \) which correspond to the probability that the \( j \) point belong to the \( k \) plan, otherwise \( \alpha_k \). The second term \( \pi(v^k_x, v^k_y, \sigma^2_x) \) is the prior density of the model parameters. If we suppose that the plans are independent, this density can be developed as the product of three densities. It just have to choose them.

Because the likelihood of the model is a gaussian, we can take for them a normal distribution \( \mathcal{N}(\mu_{prior}, \sigma^2_{prior}) \) for the parameters: \( v_x \) and \( v_y \). Those densities have to be less informative that possible. On contrary, the likelihood of the model, regarding the parameter \( \sigma^2_z \) can be write as an inverse gamma distribution (\( \mathcal{IG} \)). We thus choose for the prior distribution of this parameter a density of the same kind: \( \mathcal{IG}(\lambda_{prior}, \tau_{prior}) \), the less informative than possible.

Then by developing the equation (9), we can construct the SEM algorithm with the help of the posterior densities of the velocities \( v^k_x, v^k_y \), and the variance.
of the model error $\sigma^2\varepsilon$:

\[
p(v^k_x \mid X_{1\rightarrow N}, \tilde{X}_{1\rightarrow N}, v^k_y, \sigma^2\varepsilon) \propto N(m_{v^k_x}, \sigma^2_{v^k_x}), \tag{10}
\]

with,

\[
\sigma^2_{v^k_x} = \left( \frac{1}{\sigma^2_{\text{prior}}} + \frac{\sum_{j=1}^{N} f_{xj}^2 \tilde{X}_j^k}{\sigma^2\varepsilon} \right)^{-1} - \frac{\mu_{\text{prior}}}{\sigma^2_{\text{prior}}} - \sum_{j=1}^{N} f_{xj} (v^k_y f_{yj} + f_{tj}) \tilde{X}_j^k \right) \sigma^2\varepsilon. 
\]

For the parameter $v^k_y$, the same equation can be obtained by permuting $v^k_x$ and $v^k_y$ in equation (10).

\[
p(\sigma^2\varepsilon \mid X_{1\rightarrow N}, \tilde{X}_{1\rightarrow N}, v^k_x, v^k_y) \propto IG(\alpha, \beta), \tag{11}
\]

with,

\[
\alpha = \frac{N}{2} - 1 + \lambda_{\text{prior}} \\
\beta = \frac{1}{2} \sum_{j=1}^{N} (v^k_x f_{xj} + v^k_y f_{yj} + f_{tj})^2 \tilde{X}_j^k + \tau_{\text{prior}}.
\]

Finally, we deduce the architecture of the SEM algorithm. By having the parameters of the previous step, the first calculus is then consisting in simulating the indicator variables with a multinomial distribution: we find the probabilities for each point that it belongs to the different motion plan which are sample following the distribution of equation (8). The second calculus consists in maximizing the densities of probability of the parameters with equations (10) and (11). When the convergence of the algorithm is reach, we obtain the parameters of the different motion plans which are the velocities.
3.4 Algorithm complement

Some important points of the algorithm are interested to explain now. In natural image sequence, the spatial and temporal sampling are, in most of cases, not correctly realized (velocities are different from unity in absolute value). Then, in those image sequence, when the spatial and temporal Fourier transform is applied, the sampling creates some parallel plans to the motion plan (figure 2(b)). By misusing the language, we call them fowling up plans. An analyze of the frequency domain show the two; fowling up plans which are the nearest of the motion plan it refers and have an the most important energy are distant from the motion plan of \( d = \min(v_x, v_y, 1) \). They have thus those equation:

\[
v_x f_x + v_y f_y + f_i \pm d = 0.
\] (12)

Experimentally, we note the SEM algorithm gives some unsatisfactorily results when the fowling up plans are not taking into account. Our model is then composed of one motion plan and its two first fowling up plans of equations (12).

The architecture of SEM algorithm allow to create an rejected class. Now, we previously saw the spectrum is very noised, and despite of the threshold realized by the local maximum extraction method, a certain number of points are outliers for the model. It is thus useful for having an accurate estimation to eliminate them. For that, we use the same technic as the one use for M-estimators [9]: if the distance of the point to its plan is upper than a multiple of the error variance of the model, then the point is ejected for the calculus of
the parameters.

The stop of the algorithm is realized with the following of the error variance of the model. Indeed, all the parameters of a plan converge at the same time. This result can be observed according to the figure 4 which represents the evolution of the parameters of plans at each iterations.

![Graphs showing the evolution of parameters.](image)

Fig. 4. Example of the evolution of two plan parameters $v_x^1 = 0$, $v_y^1 = 1$, $v_x^2 = 1$ and $v_y^2 = 0$ pixel/image) for all the iterations.

The last important point is the segmentation of the cloud of points. This operation can be made in keeping at the last iteration the indicator variables of plans (figure 5). As one can see in the exemple of the figure 2(c), a plan is composed of one motion plan and two fowling up plans.
Fig. 5. Example of the separation of the three dimensional spectrum \((f_x, f_y, f_t)\) into two groups of motion plans.

4 Results

In practice, we limit the number of motion plans to two, besides no limitation in the theory prevent from having more. never mind, it is uncommon to met some transparency phenomena composed by more than two object, otherwise by more than two plans [10].

To valid the results, we have, first of all, tested image sequences composed by real images (figure 6), but whom the motion was created artificially (according to an interpolation for the non-entire velocities) in order to control the velocities and then to be able to quantify then.

In order to determine the estimation error we use the Fleet’s angular error [11]
Fig. 6. Images used to create artificially some image sequences composed of transparency phenomena. (a) Roue. (b) Caméra.

whom formal is:

\[
\alpha_e = \arccos \left( \frac{\vec{V}_{\text{est}} \cdot \vec{V}_{\text{true}}}{\|\vec{V}_{\text{est}}\| \|\vec{V}_{\text{true}}\|} \right)
\]

with \(\vec{V}_{\text{est}} = (v_{x\text{est}} \ v_{y\text{est}} \ 1)^T\) and \(\vec{V}_{\text{true}} = (v_{x\text{true}} \ v_{y\text{true}} \ 1)^T\) are the velocity vectors respectively estimated and true.

We have seen the spectrum is not ideal: face to noise, great difference between the gray level of the transparent objects. This lead to have a little number of away velocity vectors. To quantify this phenomena, we introduce a density of away which correspond to the percentage of away vectors compared with the number of expected vectors.

Another problem come from the segmentation of the image sequence in spatio-temporal windows of size \(15 \times 15 \times 15\). This integration around the analyzed pixel lead to have some velocity vectors out of the border of objects, as for all motion estimation methods which work on a neighborhood of the analyzed pixel. We thus introduce, just as an indication it, a density of surplus vectors which are evaluate as the percentage of vectors in addition compared with the
true vectors.

The first result (figure 7), represent an image sequence with transparency composed by two real images (Roue and Caméra: figure 6) move by velocity vectors respective (1, 0) and (0, 1) pixel per image. In the table 1 are shown the mean and the standard deviation (std) of the angular error all the whole optical flow calculated in each pixel of the analyzed image and the two densities.

In this ideal case of entire velocities, the results obtain are good: low mean, low standard deviation and low densities of away vectors.

Fig. 7. Image sequence 1: transparency of velocities (1,0) and (0,1) pixel/images.

(a) Analyzed image. (b) Estimated optical flow.

<table>
<thead>
<tr>
<th>velocity</th>
<th>angular error</th>
<th>density</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>mean</td>
<td>std</td>
</tr>
<tr>
<td>(1,0)</td>
<td>0.50°</td>
<td>1.68°</td>
</tr>
<tr>
<td>(0,1)</td>
<td>0.50°</td>
<td>2.58°</td>
</tr>
</tbody>
</table>

Table 1

Results corresponding to the image sequence 1.

The second result (figure 8) shows the same images (Roue and Caméra: figure 6) in transparency but with the velocity vectors of (0.2, 0.5) and (1, 0)
pixel per image. The table 2 gives the estimation results.

One can see in this more complex case the estimation is once more good but the complexity of the motion spurts back on the density of away velocity vectors.

![Image sequence 2: transparency of velocities (0.2,0.5) and (1,0) pixel/images.](image)

(a) Analyzed image. (b) Estimated optical flow.

<table>
<thead>
<tr>
<th>velocity</th>
<th>angular error</th>
<th>density</th>
</tr>
</thead>
<tbody>
<tr>
<td>(0.2,0.5)</td>
<td>3.49°, 4.28°</td>
<td>30.04%, 25.77%</td>
</tr>
<tr>
<td>(1,0)</td>
<td>1.96°, 3.73°</td>
<td>31.21%, 26.74%</td>
</tr>
</tbody>
</table>

Table 2

Results corresponding to sequence 2.

The third result (figure 9) concerns the image sequence ”Translating Tree” [12] on which we have added in additive transparency an image ”Pepsi”. The motion of the image sequence ”Translating Tree” is a translational one collinear to the horizontal axis and with an amplitude variation between 1.73 and 2.26 pixels per image. The motion of the image ”Pepsi” was generating artificially
to a velocity of (1,0) pixel per image.

We can observed the method explains allow to estimate strong velocities, around 2 pixels per image, and in manner enough accurate: means and standard deviations of the angular errors are low and this for a very low percentage of away velocity vectors.

![Fig. 9. Image sequence 3: superposition on the image sequence "Translating Tree" of an image "Pepsi" to the velocity (1,0). (a) Analyzed image. (b) Estimated optical flow.](image)

<table>
<thead>
<tr>
<th>velocity</th>
<th>angular error</th>
<th>density</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>mean</td>
<td>std</td>
</tr>
<tr>
<td>&quot;Tree&quot;</td>
<td>1.40°</td>
<td>2.09°</td>
</tr>
<tr>
<td>(1.0)</td>
<td>1.53°</td>
<td>5.08°</td>
</tr>
</tbody>
</table>

Table 3

Results corresponding to the image sequence 3.

The fourth result concerns the image sequence Yosemite [12] (figure 10). This sequence, which was creating artificially, represents the flight over a valley by a plane turning towards a point situated at right besides the horizon line. The amplitude of velocities inside this image sequence is in 0 and almost 5 pixels
The method are unable to estimate velocity higher than around 2 pixels per image, we introduce a spacial pyramid with two levels [13]. The velocities of the image sequence under-sample Yosemite are then comprise between 0 and 2.5 pixels per image. The result obtained (table 4) shows the algorithm is efficient on image sequences without transparencies. We observe nevertheless a degradation of the results for the strong motion (corner in the bottom left of the figure 10(c)).

Fig. 10. Image sequence Yosemite. (a) Analyzed image. (b) True optical flow. (c) Estimated optical flow.

Finally the fifth and the last result is concerning an image sequence entirely natural: image sequence BX (figure 11). In this image sequence, the movie
Results corresponding to the image sequence Yosemite.

camera is moving from right to left, with a motion of mean amplitude around 0.8 pixel per image and the car comes from the movie camera.

In this sequence no result can be quantify because the real optical flow is unknown. But it is interesting to note that the estimated optical flow is conform to the aspect of the motion of the movie camera and of the car. We also remarque it appeared two velocity vectors by pixel: one for each motion, in border regions of movement. Despite of the motion of the car it is not a pure translational one, the method is not trapped. So, when the non-translational motions (case of zoom like in figure 11, or a little rotation) are slow, the widening of motion plan can be approximated by the statistical model which chooses the better equivalent translational motion.

![Fig. 11. Image sequence BX. (a) Analyzed image. (b) estimated optical flow.](image-url)
5 Conclusion

The proposed method for estimating the optical flow of image sequences in taking into account the additive transparent objects in motions. it uses some hypothesis allowed to model the transparent motion, in the spatio-temporal frequency domain (Fourier transform). This model is an statistical representation of the space thanks to the SEM algorithm. The obtained results on artificial and natural image sequences, with and without transparency show the validity of such approach.

Nevertheless this method, as all frequency methods, have to face, for the moment, to a strong constraint of execution time which limits the immediate applications.

References


