Reconstructing 3D Independent Motions Using Non-Accidentalness

K. E. Ozden\textsuperscript{a} K. Cornelis\textsuperscript{a} L. Van Eycken\textsuperscript{a} L. Van Gool\textsuperscript{a,b}

\textsuperscript{a}ESAT/PSI K.U. Leuven
\textsuperscript{b}Computer Vision Lab. ETH
Leuven, Belgium Zurich, Switzerland

Abstract

Reconstructing 3D scenes with independently moving objects from uncalibrated monocular image sequences still poses serious challenges. One important problem is to find the relative scales between these different reconstructed objects. The perspective reconstruction of a single object can only be known up to a certain scale which results in a one-parameter family of relative scales and trajectories per moving object. The paper formulates this ambiguity and proposes solutions in the vein of "non-accidentalness". Two instantiations of its use are analyzed: planar motion and the 'heading constraint'.

1 Introduction

Structure-from-Motion (SfM) has been a popular research topic in computer vision for at least two decades now. Most of this work is restricted to static scenes [7]. A typical scene naturally contains independently moving objects however.

There has recently been a surge of papers on the detection and analysis of reconstructing dynamic scenes. With their multibody factorization method, Costeria and Kanade [4] have extended the static scene factorization method of Tomasi and Kanade [13] to scenes with multiple, independently moving rigid objects. Fitzgibbon and Zisserman [5] demonstrated the usefulness of multiple, independently moving rigid objects in self-calibration. Machline \textit{et al.} [9] used linear subspace constraints to segment a dynamic scene into its independently moving, but not necessarily rigid, components. Wolf and Shashua [18] introduced the two-body algebraic structure called segmentation matrix which was extended by Vidal \textit{et al.} [15] to multiple moving objects. Brand [2] and Bregler \textit{et al.}'s work [3] showed that 3D reconstruction is possible even when the object is moving non-rigidly. There is also interesting research where the moving object is supposed to move in a constrained way. Avidan and Shashua [1] investigated the case where a point is moving on a line or a conic. Han and Kanade [6] demonstrated the calibration of a single camera when objects are moving on a line with constant speed. Sturm [12] analyzed the case of a stereo camera observing points that are moving on a plane. Wolf and Shashua [17] extended the classical notion of three- to two-dimensional image projection to n-dimensional space for non-rigid scenes.

In this paper, we follow a path similar to Avidan and Shashua [1] where a single point is moving on a line or a conic. However, our constraints are geometrically less restrictive, at the expense of requiring full projection matrices for the object reconstructions. We will assume that segmentation has already been performed (several techniques have been proposed to achieve this [4, 9, 15, 14, 11]). Rather than imposing strict geometric assumptions, we propose a \textit{non-accidentalness} principle: among all possible relative scales, choose the one for which the reconstruction exhibits some properties which are unlikely to be encountered accidentally. The non-accidentalness principle has been introduced and successfully exploited first in the area of grouping and object recognition [8].

3D reconstruction from uncalibrated images has an inherent scale ambiguity [7]. When there are multiple moving objects in a scene, the relative scale ambiguity between their reconstructions will result in an unrealistic total reconstruction. A one-parameter family of ambiguity exists for every object with respect to the static background. Consider a video of a moving car. Without giving high level knowledge about the world, a computer algorithm cannot distinguish between a small toy car flying in front of the camera or a real car further away on the road. Since all these reconstructions are equally correct in terms of geometry, the only way to solve this problem is to introduce further assumptions on the scales and/or the trajectories. In this paper, we propose two such constraints, as examples of the much wider principle of non-accidentalness.

The paper is organized as follows. Section 2 discusses the relative scale problem in more detail. In this section we derive the basic equation on which the non-accidentalness principle is based. Section 3 discusses this principle and analyzes two such cases which are planar motion and the 'heading constraint'. The viability of these ideas is corrob-
orated through real life experiments. Section 4 summarizes the main ideas of the paper and discusses future work.

2 The Relative Scale Problem

As mentioned before, the relative scale ambiguity is caused by the unknown scale factor in the reconstruction of a single object. To analyze this phenomenon, consider an image sequence of a scene which is static, except for one rigid, independently moving object. This restriction on the number of moving objects is introduced to simplify the discussion. The case of several independently moving objects can be dealt with similarly. For every frame $i$ of moving objects is introduced to simplify the discussion.

As mentioned before, the relative scale ambiguity is caused to the moving object. The static part of the scene — 'the background' — and with respect to the moving object. The $3 \times 3$ rotation matrices $R_o^i$ and $R_x^i$ represent these two orientations respectively. The $3 \times 1$ translation vectors $t_o^i$ and $t_c^i$ represent these positions. What we would like to find is the rotation $R_o^i$ and the translation $t_o^i$ which represent the motion of the object in the scene for every frame $i$. These transformations and their corresponding notation are illustrated in Fig. 1. The relation among them can be written as follows:

$$
\begin{bmatrix}
R_x^T \\
0
\end{bmatrix} =
\begin{bmatrix}
R_x^T & -R_x^T t_c^i \\
0 & 1
\end{bmatrix}
\begin{bmatrix}
R_o & t_o^i \\
0 & 1
\end{bmatrix}
$$

where we have dropped the frame indices for compact notation. $T$ stands for the matrix transpose, $I$ is a $3 \times 3$ identity matrix and $0$ is a $1 \times 3$ vector of zeroes. The left hand side states that, to directly transform a point from the object coordinate system to the current camera coordinate system, apply the relative transformation defined by $R_o^i$ and $-R_x^T t_c^i$. The right hand side of this expression tells us that, alternatively, to transform a point in the object local coordinate system to the current camera coordinate system, first rotate it around the object coordinate system with $R_o^i$, perform the actual translation $t_o^i$, and as a last step transform it to the current camera coordinate system using $R_x^T$ and $-R_x^T t_c^i$. The rotation and translation parts of Eq. (1) yield

$$
R_x^T = R_x^T R_o 
$$

$$
- R_x^T t_c^i = R_x^T t_o^i - R_x^T t_c^i 
$$

If we fix the coordinate system of the reconstructions to an arbitrary location, the rotations $R_x^T$ and $R_o^T$ are unambiguous. So we can find exact $R_x^T$ matrices using Eq. (2). However, due to the nature of uncalibrated perspective SfM, we cannot extract the translation of the camera with respect to the background and the object at an absolute scale [16]. Since only the relative scale is important for us, we can fix the scale for the background. However, there still remains the relative scale of the moving objects. Given the fact that $t_o^i$ is only known up to a relative scale $s$, Eq. (3) will give a different $t_o^i$ for every scale.

$$
s(-R_x^T t_o^i) = R_x^T t_o^i - R_x^T t_c^i 
$$

where $t_o^i$ is the object trajectory at scale $s$. Suppose $t_o^i$ represents the true object trajectory. Merging equations (3) and (4), and multiplying both sides with $R_x$ yields the following relation between the true trajectory of the object $t_o$ and the computed trajectory $t_o^i$ with an incorrect, relative scale factor $s \neq 1$.

$$
t_o^i = s t_o^i + (1 - s)t_c^i 
$$

Hence, the object translation $t_o^i$ which one finds at the relative scale $s$ is a linear interpolation of the true object translation $t_o$ and the camera translation $t_c$. This leads to a one-parameter family of object trajectories $t_o^i$ for different relative scale values $s$, all equally compatible with the image data. When $s = 1$, $t_o^i$ equals the true $t_o$ and when $s$ gets closer to zero, $t_o^i$ evolves towards the camera path. For different values of $s$ other than 1, $t_o^i$ will always undergo the influence of the camera motion $t_c$.

A further analysis shows that such a coupling exists not only for the translation components of the object motion (the path of the origin of the object coordinate system) but also for any point attached to the object coordinate system. Assume $p^0$ is the position of a point on the object in the first frame. Its position $p$ in frame $i$ can be written as:

$$
p = R_i p^0 + t_o^i 
$$

in which the frame index $i$ is dropped again to simplify the notation. Similarly, its scaled version $p_s$ moves according to the following transformation:

$$
p_s = R_s p_s^0 + t_o^i 
$$

Figure 1: Transformations between the static and the dynamic parts of the scene.
Without loss of generality, we can assume the fixed world coordinate system to be attached to the initial camera pose (this is also assumed for the rest of the paper). Then, \( p_0 \) is equal to \( sp_0 \). Introducing this fact and Eq. (5) into the above equation yields:

\[
p_s = s(Rp^0 + t_o) + (1 - s)t_c \tag{8}
\]

and by incorporating Eq. (6), it can be written as

\[
p_s = sp + (1 - s)t_c \tag{9}
\]

which is similar to Eq. (5).

However, the input coming from the uncalibrated SfM software is not \( t_o \) or \( p \) but some \( t_o \) or \( p_s \) (as defined by eqs. (5) and (9)). So one should reverse the procedure, to find the true object trajectory and corresponding relative scale. If one brings \( p \) to the left side of Eq. (9), it becomes:

\[
p = mp_s + (1 - m)t_c \tag{10}
\]

where \( m = 1/s \). A similar expression can also be written for Eq. (5). The correct \( m \) would remove the additive components of \( t_c \) in \( p_s \) or \( t_o \). To find such \( m \) is the topic of the next section.

3 The non-accidentalness principle

It is clear from Eq. (5) that while computing the object trajectories, there is a systematic addition of the camera translation at wrong relative scales. Consequently, we expect the computed object motion to lose special properties, such as planarity, periodicity, piecewise linearity, constant speed, etc., at wrong scales. This is the basis of our non-accidentalness principle. To demonstrate it, consider the robot arm in Fig. 2. Its trajectory consists of three straight line segments which are perpendicular to each other. Fig. 3 shows the reconstructed ball trajectory for different relative scales including the correct one. As we get further away from the correct scale, the trajectory loses its simple shape. In this paper, only two of the many possible non-accidental properties will be explored: the ‘planarity constraint’ and the ‘heading constraint’.

3.1 The planarity constraint

In this section, the existence of a planar trajectory among the one-parameter family of relative scales, is taken as an indication of non-accidentalness.

The usefulness of the planarity constraint depends on the coupled nature of the object and the camera trajectory. A degenerate case occurs when the camera moves parallel to the plane of the object motion or when there exists an affine

Figure 2: Several frames from the robot sequence. The robot translates the ball through the static scene. The hand-held camera moves around while taking the images.

Figure 3: The ball trajectory as it is found from uncalibrated SfM for different relative scales, and between which it cannot decide. The applied relative scale factors \( m \) are 0.20, 0.25, 0.30, 0.36, 0.40, 0.45 from left to right. 0.36 is the correct relative scale (middle right).
transformation between concurrent camera and object positions. A more extensive discussion of these degenerate cases is given elsewhere [10].

There are several techniques to impose the planarity criterion among the one-parameter family of possible object trajectories. Principal Component Analysis (PCA) is one way to achieve it. From equations (2), (3) and (6) we find the trajectory of a point on the object as:

$$ p^i = t^i_s + R^i_o m p^0_s - R^i_o m t^0_s $$

(11)

where $p^0_s = s p^0, t^0_s = s t^0_s$, and $m = 1/s$. Such substitutions are necessary since we only know some scaled version of these vectors. We are looking for the scale value which makes the left hand side the most planar. As a planarity criterion, we look for the minimum of the determinant of the corresponding ‘covariance matrix’ computed from the 3 components of the object trajectory $t_o$. The determinant of this matrix indicates the volume of the space spanned by the principal components which should be zero in the case of a planar trajectory. However due to the effect of noise, this determinant will never be zero so we prefer the minimization of this determinant over finding its roots directly.

First, we translate both sides of Eq. (11) to their mean. We get an expression of the form $m A^i + B^i$ in which $A^i$ represents the mean-shifted $R^i_o p^0_s - R^i_o t^0_s$ and $B^i$ represents the mean-shifted $t^i_s$. The expression we want to minimize is:

$$ V = det \left( \sum_{i=0}^{\#points} (m A^i + B^i)(m A^i + B^i)^T \right) $$

(12)

This is a polynomial of $6^{th}$ degree in $m$. To minimize it, one can take the derivative, and find the roots with numerical methods. Among the roots the one which gives the minimal value is chosen.

If the object only translates on a plane and rotates around an axis orthogonal to that plane, i.e. planar motion [7], the above solution always works since every point on the object moves on a plane. But we are interested in more general cases where only a part of the object is moving planarly. As an example, take a ball that is rolling on a planar surface. None of its observable points is performing a purely planar motion, although the global trajectory of the ball is planar. Only the centroid is moving planarly. The average of the observable points will not coincide with the real centroid. To find the real centroid and the relative scale, we propose an iterative approach where the point cloud centroid and the PCA solution for $m$ are used as initial values. Starting from these data, a consecutive refinement step tries to simultaneously come up with an enhanced relative scale and a point with maximally planar motion for that scale. Hence, we have to solve for four unknowns, namely the values for the three coordinates of $p^0_s$ – the centroid in this case – and $m$. We used Levenberg-Marquardt (LM) to minimize the ratio of the third to the second eigenvalue of the covariance matrix in order to find the point that yields the maximal planarity.

To demonstrate the usefulness of this approach we experimented with video data of a ball rolling on a ground plane. Fig. 5 demonstrates the results. Two possible trajectories of the ball and their reconstructions are shown in front of the reconstructed background. The ball at the wrong scale is indicated by a square around it, and the correctly scaled one is shown with a triangle. As can be seen from the bottom image of Fig. 5, the latter trajectory is close to planar and parallel to the ground plane. The other trajectory is not planar at all and the ball is found flying in the air.

Figure 4: Several frames from the original ball sequence which is 400 frames long.

### 3.2 The heading constraint

Many types of moving objects, such as humans, cars, bikes etc. have a natural frontal side and therefore heading direction. Hence, these heading directions or vectors are usually parallel to the tangent of the object trajectory. If not, the objects would undergo strange motions like cars going into a skid. This does not usually happen. The mathematical equation describing this ‘heading constraint’ is:

$$ l^{ij} R^{ij}_o v^i_o = v^j_o $$

(13)

where $R^{ij}_o$ is the rotation of the object from frame $i$ to frame $j$. $l^{ij}$ is a scale factor. $v^i_o$ is the tangent to the object’s trajectory at frame $i$ which can be approximated by:

$$ v^i_o = g^{i+1}_o - g^{i-1}_o $$

(14)
changing all subscripts

Equations (14) and (13) together with a relative scale constraint can not help us. In such cases, Eq. (13) will hold ever, also here, there are degenerate cases where this constraint tend to lead to the violation of the heading constraint. How-
scales. Just as they may destroy planarity, they would also

Camera velocities.

Similar expression is also used for the approximation of the video sequences with their relatively high frame rates. A frame. This is a valid approximation since we generally use

Figure 5: The reconstructed scene for the ball sequence from different views. The rectangle denotes the un-scaled ball reconstruction and the triangle denotes the scaled version of the ball reconstruction.

where \( \mathbf{g}_i \) is the position of the centroid of the object at \( i \)th frame. This is a valid approximation since we generally use video sequences with their relatively high frame rates. A similar expression is also used for the approximation of the camera velocities.

Eq. (5) states that the object trajectory will contain components from the camera translation for the wrong relative scales. Just as they may destroy planarity, they would also tend to lead to the violation of the heading constraint. However, also here, there are degenerate cases where this constraint can not help us. In such cases, Eq. (13) will hold for every relative scale. To analyze such cases, let us apply equations (14) and (13) together with a relative scale \( s \). By changing all subscripts \( o \) to \( os \), the following equation results if the heading constraint is to hold for other, incorrect scales \( s \):

\[
k \mathbf{R}^{ij}_{os} (\mathbf{g}_\text{os}_{i+1} - \mathbf{g}_\text{os}_{i}) = \mathbf{g}_\text{os}_{i+1} - \mathbf{g}_\text{os}_{i-1}
\]

where \( k \) is a scale factor. Then we introduce Eq. (9) to come up with the following equation:

\[
k \mathbf{p} = \mathbf{q}
\]

where

\[
\mathbf{p} = \mathbf{R}^{ij}_{os} (s(\mathbf{g}_o^{i+1} - \mathbf{g}_o^{i-1}) + (1 - s)(\mathbf{t}_c^{i+1} - \mathbf{t}_c^{i-1})) \\
\mathbf{q} = s(\mathbf{g}_o^{i+1} - \mathbf{g}_o^{i-1}) + (1 - s)(\mathbf{t}_c^{i+1} - \mathbf{t}_c^{i-1})
\]

By introducing \( \mathbf{v}_o^i, \mathbf{v}_c^j, \mathbf{v}_o^j, \mathbf{v}_c^j \), Eq. (16) turns into:

\[
k(s \mathbf{R}^{ij}_{o} \mathbf{v}_o^i + (1 - s) \mathbf{R}^{ij}_{c} \mathbf{v}_c^j) = s \mathbf{v}_o^i + (1 - s) \mathbf{v}_c^j
\]

If the above equation holds for values of \( s \) other than 1, one has a degenerate case. If we insert Eq. (13) into the above equation and leave the term \( \mathbf{R}^{ij}_{o} \mathbf{v}_o^i \) on the left side, we come up with

\[
\frac{s}{1 - s} (k - l) \mathbf{R}^{ij}_{o} \mathbf{v}_o^i = \mathbf{v}_c^j - k \mathbf{R}^{ij}_{c} \mathbf{v}_c^j
\]

which should still hold for different choices of \( i \) and \( j \). The left hand side spans a line passing through the origin for different values of \( s \). This results in a constraint on the right hand side that \( \mathbf{v}_c^j \) is a constant multiple of \( \mathbf{R}^{ij}_{c} \mathbf{v}_c^j \). For a degenerate case to occur, this constraint should hold for every frame pair which means that all the camera translations should be determined by the object rotation. Fortunately, this is really hard to find in real life except for some simple motion cases. One such example is the case where both the camera and the object move on a line and the directions of these lines are in keeping with Eq. (18).

Let us return to the actual use of the constraint. Given two frames, finding the relative scale amounts to solving a polynomial equation which is formulated next. Merging Eq. (10) and Eq. (14) yields:

\[
\mathbf{v}_o^i = m \mathbf{v}_o^j + (1 - m) \mathbf{v}_c^j
\]

The expression we want to maximize is coming from the heading constraint in Eq. (13). For the corresponding parallelism of velocity vectors to hold, we can maximize the cos between the vectors:

\[
\mathbf{a}^i \mathbf{b} = \cos(\mathbf{R}^{ij}_{c} \mathbf{v}_c^j, \mathbf{v}_o^i)
\]

where

\[
\mathbf{a} = \frac{\mathbf{R}^{ij}_{c}(m \mathbf{v}_o^j + (1 - m) \mathbf{v}_c^j)}{\sqrt{(m \mathbf{v}_o^j + (1 - m) \mathbf{v}_c^j)^T (m \mathbf{v}_o^j + (1 - m) \mathbf{v}_c^j)}}
\]

\[
\mathbf{b} = \frac{m \mathbf{v}_o^j + (1 - m) \mathbf{v}_c^j}{\sqrt{(m \mathbf{v}_o^j + (1 - m) \mathbf{v}_c^j)^T (m \mathbf{v}_o^j + (1 - m) \mathbf{v}_c^j)}}
\]
This is the scalar product of two normalized vectors and it has the form of a rational polynomial. One can maximize the square of the cosine expression in Eq. (20) in case of an image sequence where the object suddenly decides to go backwards somewhere in the sequence. We discarded such rare cases to simplify the solution.

Solving for \( \mathbf{m} \) with different frames \( i \) and \( j \) results in different \( \mathbf{m} \)'s. One reason is the fact that an object may not always follow its heading perfectly. For example a person may twist his torso for a few frames. Such cases should be treated as outliers and we can use Eq. (20) in a RANSAC [7] scheme for a robust estimation of \( \mathbf{m} \). Therefore, several random choices of \( i \) and \( j \) were made, and the \( \mathbf{m} \) with maximal support was chosen, i.e. depending on how many other \( i, j \) pairs have a super-threshold value for that \( \mathbf{m} \) according to Eq. (20).

Another problem we saw in our experiments is that ‘objects’ like humans obey the heading constraint globally but not instantaneously, e.g. during a single step. The center of gravity of the torso oscillates between left and right at this level of granularity. To avoid that, while calculating velocities, especially for human gait, we suggest to use the formula

\[
\mathbf{v}^i = t_i + n - t_i - n
\]

where \( n \) depends on the speed of the person and the sampling rate of video. We can estimate a good value for \( n \) during RANSAC random sampling, as an additional parameter to be estimated.

At the end, an additional refinement step is included on the \( \mathbf{m} \) value supplied by RANSAC. The selection of the optimal \( \mathbf{m} \) is based on the values for which the inliers minimize an error functional of the form

\[
\sum_{\text{inliers}} \angle \mathbf{v}^i \mathbf{v}^o \mathbf{v}^j
\]

around the initial value of \( \mathbf{m} \). The reason for using angles rather than cosines as in Eq. (20) is the fact that angles are geometrically more meaningful.

Figure 6: Three frames from the original walking sequence consisting of 61 images.

To demonstrate the usefulness of our algorithm, we recorded the video of a person walking in front of a static scene. Several frames from the original 61 frame sequence are shown in Fig. 6. We reconstructed the background and the person separately using our sequential perspective SFM software. We were only able to reconstruct the upper part of the body since the lower parts are non-rigid and do not have enough features. The optimal \( n \) we found is 8, which is approximately the duration of a single step of the person. The results are shown in Fig. 7. Every row corresponds to a different virtual camera position. They depict the reconstructed 3D world at the time that the middle frame of Fig. 6 was shot. The right column corresponds to the reconstruction at an incorrect relative scale. The circle surrounds the tiny human at the wrong scale. The left column shows the result after we applied the relative scale we found using the heading constraint. Note that for the cam-

Figure 7: Reconstruction of the walking sequence in Fig. 6 from different virtual camera positions. The right and the left column depict, resp., the reconstruction applying and incorrect and the correct relative scale factor, with the latter found using the heading constraint. The ellipses in the right column encircle the tiny reconstructed person, obviously at the wrong scale.
era positions corresponding to the original images, both left and right images will look the same. As can be seen, the relative scale we found with the heading constraint gives quite realistic results. To further corroborate the usefulness of the non-accidentalness assumption, we applied the planarity constraint to the same sequence and came up with a close solution. They differed less than 3 percent in terms of scale [10].

4 Conclusion

Reconstructing scenes containing independently moving objects remains a big challenge for the computer vision community. One problem which did not receive much attention is the relative scale ambiguity between the reconstructions of different, independently moving parts of the scene. In this paper, we gave a simple formulation for the problem and demonstrated that at wrong relative scales, object trajectories would contain components of the camera translation. If the object motion has some special property, such as planarity, periodicity etc., we expect those properties to be lost at wrong relative scales due to the effect of camera translation. This observation forms the basis of our “non-accidentalness” principle which states that among several mathematically correct solutions, if one of them results in an object motion with special properties, this is not accidental and therefore, it is very likely that we are at the correct scale. We demonstrated the use of two such properties. One of them is the assumption of planar motion since many objects move on a plane. In this approach, we simultaneously search for a point on the object and the scale which makes this point’s motion planar. As a second property, we investigated the heading constraint which is based on the fact that objects generally move in their heading direction. We demonstrated the applicability of both techniques with real life experiments.

Future work will focus on the automatic selection and possible combination of multiple non-accidental properties. Even if a single non-accidentalness criterion may not work, eg. when a camera travels in a line parallel to that of the action, a combination of several like the two proposed might. So, in the end the system would need to explore a battery of special properties and arbitrate.

Finally, it is worthwhile mentioning that Eq. (5) gives rise to a second scale selection criterion, the ‘independence criterion’, as an alternative to non-accidentalness. More on this can again be found in [10].

Acknowledgement: The authors gratefully acknowledge support by the Flemish Fund for Scientific Research and the K.U.Leuven Research Council GOA project ‘MARVEL’.

References