Guided Quasi-Dense Tracking for 3D Reconstruction

Andrei Khropov*
Laboratory of Computational Methods, MM MSU

Anton Konushin†
Graphics & Media Lab, CMC MSU

Abstract

In this paper we propose a framework for obtaining quasi-dense Euclidean structure reconstruction by means of guided quasi-dense point tracking in image sequences. We use a stratified algorithm that establishes most reliable sparse point correspondences first then robustly estimates multiview geometry and finally propagates sparse point features to quasi-dense matches and tracks. Notable properties of our algorithm are built-in motion model selection using geometric robust information criteria (GRIC) that helps to avoid common algorithm degeneracy. Matching outliers are segmented by robust homography estimation for rotational camera movement or by reprojection error thresholding otherwise. We employ automatic keyframe selection that is used to provide frames with reasonable disparity for reliable matching and reconstruction. These technique also reduces cost of computationally expensive quasi-dense tracking.

Keywords: quasi-dense matching, quasi-dense tracking, point tracking, 3D reconstruction, structure from motion

1 Introduction

The problem of automatic 3D reconstruction from image sequences or structure-from-motion is one of the key areas of research in Computer Vision as it is important from both theoretical and practical points of view. An enormous progress has been made in this area during the last decade and theoretical developments have reached the level of maturity for a textbook [Hartley and Zisserman 2004]. The most successful approach to uncalibrated 3d reconstruction is based on establishing correspondences on multiple images between so called image features (distinguishing objects with geometrical properties). The simplest and most widely used kind of feature is a point. Traditionally, there were two kinds of point correspondences between images: sparse and dense. Sparse correspondences are established between distinct spots in the images. They can be detected and matched with subpixel precision and therefore are reliable for geometry estimation. In contrast, dense correspondence are established for every pixel of the images. Because generally each pixel cannot always be reliably distinguished from other pixels especially in low-textured regions unconstrained dense matching remain an intractable problem. All existing dense matching methods relies on relative camera positions, known a priori.In addition this approach is very expensive computationally, so it is usually limited to several images with small camera motion and calibrated cameras setup. In short, sparse points are precise but there are too few of them to reconstruct shape of scene objects. Dense correspondences could possibly allow full scene depth recovery but are error-prone. The dense-based approach usually starts with sparse matching to obtain information for camera calibration estimation, and then use this information for 3d reconstruction via dense matching.

General pipeline for structure and motion estimation from image sequences proceeds as follows:

- Establish point correspondences between consecutive frames. Robustly estimate projective 2-view and 3-view geometry encoded in fundamental matrix and trifocal tensor respectively. Inconsistent matches are detected and discarded at this stage;
- Projective reconstruction for the whole sequence is obtained by merging of two- and three-view reconstructions for subsequences;
- Projective reconstruction is upgraded to metric using self-calibration;
- Finally, accuracy of reconstruction is refined by means of point reprojection error minimization w.r.t. points' positions in 3d and camera parameters. This process is called *bundle adjustment*.

In 2002 a new quasi-dense approach for establishing correspondences was proposed by Lhuillier and Quan [2002a] [2002b] which is essentially a compromise between the aforesaid two. Quasi-dense points retain precision and reliability of sparse matches, while greatly surpassing them in sheer numbers and uniformity of coverage of scene surfaces. This idea is crucial to our work and discussed in depth below.

2 Related work and discussion

Two general approaches exist for correspondence estimation. The first is called feature matching and consists of two steps - independent detection of features in all frames and their matching in certain frame pairs (usually successive ones). The second is called feature tracking. It proceeds by sequential tracking of positions of once detected features, usually by searching for a single best candidate. On of the most popular algorithm of this kind is Lucas-Kanade-Tomasi tracker [1991]. Another popular approach performs an exhaustive search through possible matches measuring their similarity by some cost function. SSD (Sum of Square Differences) or cross-correlation are usually employed. The latter is preferable since it is more robust to changes in image brightness and therefore used in most modern algorithms. Search space is usually bounded either by assumption of small frame-to-frame displacements which is the case in video sequences or restricted by motion model and/or already estimated projective geometry. The latter case is referred to as *guided matching/tracking*. In this work it is used for establishing quasi-dense correspondences.

Algorithms that employ cross-correlation matching or tracking combined by robust fundamental matrix and trifocal tensor estimation are an established standard in sparse tracking [Pollefeys et al. 1998], [Fitzgibbon and Zisserman 1998], [Pollefeys and Gool 2002]. Methods that are used to improve the performance of such algorithms by selecting motion model and using the concept of keyframes can be found in [Gibson et al. 2002], [Thormählen et al.

*e-mail: akhropov@fit.com.ru
†e-mail:ktosh@graphics.cs.msu.ru

2004], [Konushin et al. 2005]. In our work this issue is addressed too.

Methods for obtaining euclidean reconstruction for extended image sequences are discussed in [Hartley and Zisserman 2004], [Fitzgibbon and Zisserman 1998].

As it has been already noted in the introduction section sparse correspondences provide too sparsely scattered 3d points as a reconstruction that are not able to represent shape of scene objects that is why additional techniques are necessary. Dense matching is discussed in [Scharstein and Szeliski 2002]. But dense methods are not always reliable as weakly textured areas and very computationally expensive.

Lhuillier and Quan [2002a] proposed a method they called *quasi-dense matching* that is positioned as the golden mean between sparse and dense approaches. In a quasi-dense process frame-to-frame quasi-dense matches are evenly distributed within textured areas of images. This property leads to even distribution of corresponding reconstructed 3d points and makes shape reconstruction possible [Zeng et al. 2004],[Lhuillier and Quan 2005]. Their approach to constructing multi-frame quasi-dense point tracks is presented in [Lhuillier and Quan 2002b]. We propose a guided quasi-dense tracking that is described in section 4. It differs from the method presented in [Lhuillier and Quan 2002b] in several ways. It uses most reliable sparse correspondences to estimate geometry first. Sparse seed points are uniformly distributed to minimize possibility of degenerate configurations [Konushin et al. 2005]. Resampled quasi-dense points are selected as most distinct points in local blocks instead of just being their centers. We perform guided quasi-dense tracking instead of matching to maximize number of views where each feature is present. Our approach also incorporates selection of keyframes and motion model selection [Gibson et al. 2002], [Thormählen et al. 2004], [Konushin et al. 2005] for improved reliability.

The process of obtaining reconstruction of euclidean 3d structure from point matches is described in detail in [Hartley and Zisserman 2004] and [Pollefeys et al. 1999]. The final step of this process is bundle adjustment which is usually performed using sparse Levenberg-Marquardt algorithm (see [Hartley and Zisserman 2004] and [Triggs et al. 1999] for details). We use this technique as a final step too.

Algorithm pipeline outline is presented in Figure 1.

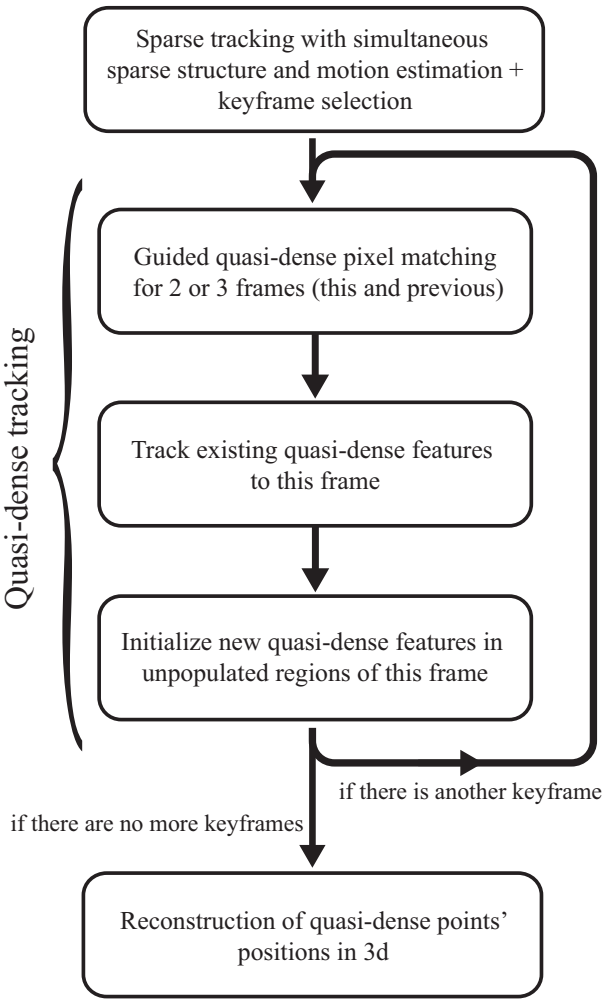

Figure 1: Algorithm pipeline outline

3 Sparse tracking and camera calibration

Quasi-dense matching methods are based on propagation of point correspondences in neighborhood of seed matches supplied elsewhere. In the most basic case, seed correspondences can be specified manually. However, several factors should be taken into account during seed matches generation. First, because quasi-dense matches are established in an iterative fashion by propagating seed matches (see Section 4 for details) absence of seed points in certain areas forbids creation of quasi-dense matches in these areas and unreliable (outlier) seed points either completely cease propagation or provide erroneous quasi-dense pixel matches in their neighborhoods. Second, because quasi-dense points are less discriminative than general point features, the tracking of quasi-dense points is subject to greater error build-up and reliability of quasi-dense tracks should be checked via precise geometrical constraints. Third, establishing of quasi-dense correspondences and robust estimation of multi-view relations from quasi-dense matches is very computationally expensive task.

These factors place additional requirements on seed matches generation. To consider all of them we adapt a sparse feature tracking method described in [Konushin et al. 2005] for seed matches generation. This method is based on partitioning the input image sequence into several segments by adaptive selection of a number of keyframes. Keyframes are selected iteratively one-by-one so that either homography or fundamental matrix is reliably estimated from their matches while preserving the sufficient number of sparse point tracks. We apply quasi-dense matching to keyframes only. Such frames have sufficient number of sparse matches for seed points for reliable quasi-dense matching and corresponding viewpoints are relatively distant from each other to provide higher than for all frames precision of estimation of 2d points positions. This technique also lowers the computation cost of quasi-dense tracking. Adaptive key-frame selection with correct corresponding two-view relation selection via GRIC [Konushin et al. 2005] allows correct guidance for quasi-dense matching under arbitrary camera movement.

To increase image area coverage and provide sufficient number of precise seed points our sparse tracking algorithm uniformly selects features in images. Each image is partitioned into set of rectangular regions which are called bins. Let N be the number of bins and M is desired number of features. M/N best features are selected from each bin and used for tracking (see [Konushin et al. 2005] for

details). The idea is illustrated in Figure 2.

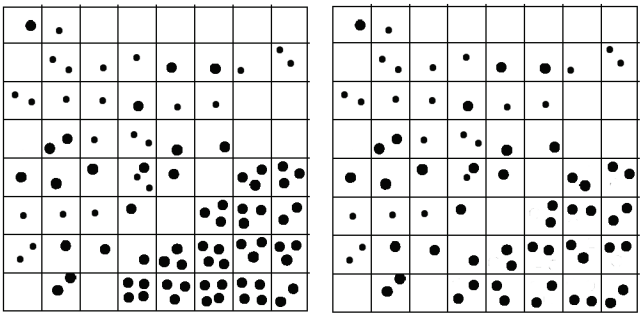

Figure 2: Partitioning features into bins. Size of circle represents relative feature quality. Left: detected sparse features. Right: filtered sparse features, some good points have been neglected in densely populated bins

Described tracking algorithm includes computing 2-view (fundamental matrix or homography) and 3-view (trifocal tensor) geometric constraints that allows us to perform guided matching and tracking of quasi-dense features.

As a final step sparse structure and motion are recovered using algorithm similar to [Fitzgibbon and Zisserman 1998]. Recovered cameras permit quasi-dense feature track filtration by reprojection, see Section 4.3.

4 Establishing quasi-dense matches

In this section three main algorithms are described in detail. The first two are variations of quasi-dense matching for 2 and 3 images respectively. They are employed by tracking procedure which is described in the third section as a means of establishing putative correspondences between successive pairs and triples of keyframes. Overall tracking pipeline is presented in Figure 1.

4.1 Quasi-dense matching for 2 frames

Pixel-to-pixel quasi-dense correspondences are obtained at first. This process is generally the same as in [Lhuillier and Quan 2002a]. They are established through recursive propagation of existing pixel-to-pixel matches. This process is initialized by seed matches that are sparse matches downsampled to pixel precision.

Match quality is measured by Zero-mean Normalized Cross-Correlation (ZNCC) that is invariant to slight brightness variations:

$$\frac{\sum_{\Delta\in W_r}(I^1(\mathbf{x}_1+\Delta)-\overline{I^1}(\mathbf{x}_1))\cdot(I^2(\mathbf{x}_2+\Delta)-\overline{I^2}(\mathbf{x}_2))}{\sum_{\Delta\in W_r}\left(I^1(\mathbf{x}_1+\Delta)-\overline{I^1}(\mathbf{x}_1)\right)^2\cdot\sum_{\Delta\in W_r}\left(I^2(\mathbf{x}_2+\Delta)-\overline{I^2}(\mathbf{x}_2)\right)^2}$$

where $\mathbf{x}_1=(x_1,y_1)^T$, $\mathbf{x}_2=(x_2,y_2)^T$ – coordinates of compared pixels, $I^i(\mathbf{x})$ – intensity of pixel $\mathbf{x}$ in ith image, $W_r=\{(i,j)|i,j\in\overline{-r,r},\ r\in\mathbb{N}\}$ – correlation window with radius r (typically 5), $\overline{I}(\mathbf{x})$ – mean intensity in window W with center $\mathbf{x}$. ZNCC $\in(0,1)$. Only matches with ZNCC greater than a certain threshold z are chosen as possible candidates. This threshold depends on correlation window. For quasi-dense matching we use relatively small window with $r=2$ and non-restrictive ZNCC threshold of 0.5.

To be selected a quasi-dense pixel match must also pass *cross-consistency check*, i.e. $ZNCC(\mathbf{x},\mathbf{y})$ must be maximum among possible candidate matches for $\mathbf{x}$ in the second image as well as for $\mathbf{y}$ in the first image. To limit a variety of possible candidate matches *2d disparity gradient limit* and epipolar constraint or homography (2-frame specific) are used.

2d disparity gradient limit assumes optical flow smoothness and limits possible candidates to close neighbors of already established match $(\mathbf{x}^1,\mathbf{x}^2)$ being propagated (see pseudocode below). If $\mathcal{N}(\mathbf{x})=\{\mathbf{x}+\Delta|\Delta\in W_n\}$ where W_n is a neighborhood window (n is typically 2) then 2d disparity gradient limit restricts new possible matches to

$$\mathcal{N}(\mathbf{x}^1,\mathbf{x}^2)=$$
$$\{(\mathbf{u}^1,\mathbf{u}^2)|\mathbf{u}^1\in\mathcal{N}(\mathbf{x}^1),\ \mathbf{u}^2\in\mathcal{N}(\mathbf{x}^2),\ \|(\mathbf{u}^1-\mathbf{u}^2)-(\mathbf{x}^1-\mathbf{x}^2)\|_\infty\le d\}$$

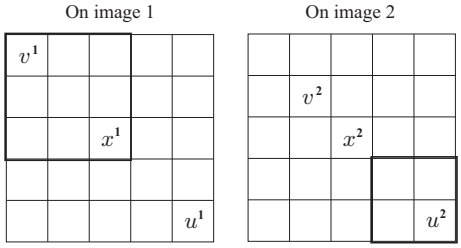

Figure 3: Neighborhood of $(\mathbf{x}^1,\mathbf{x}^2)$ where possible matches for $\mathbf{u}^1$ are in the frame in the neighborhood of $\mathbf{x}^2$ and possible matches for $\mathbf{v}^2$ are in the frame in the neighborhood of $\mathbf{x}^1$

d is chosen to be the smallest possible (1) to limit bad matches at the occluding contours in [Lhuillier and Quan 2002a]. We relax this restriction (make it 2) since bad matches are filtered out by epipolar and trifocal constraints.

Confidence measure $s(\mathbf{x})$ is used to select pixels that can be chosen as a quasi-dense pixel feature. To prohibit propagation is weakly textured regions it is simply maximum intensity difference with the closest neighbors:

$$s(\mathbf{x})=\max_{\Delta\in W_c}|I(\mathbf{x}+\Delta)-I(\mathbf{x})|$$

c is usually 1. $\mathbf{x}$ with $s(\mathbf{x})$ less than a threshold t (typically 1-2% of maximum image intensity) are rejected.

The propagation algorithm pseudocode is presented in Algorithm 1.

After quasi-dense pixel matches have been computed we obtain local homographies and resampled quasi-dense subpixel correspondences. We subdivide first image into rectangular blocks and robustly (using RANSAC) fit a local affine transformation H to pixel quasi-dense correspondences $(\mathbf{u}^1,\mathbf{u}^2)$ whose first point $\mathbf{u}^1$ is within the block assuming that most pixel matches within the block are approximately lying on the single planar patch [Lhuillier and Quan 2005]. RANSAC also provides information about pixel matches that do not fit to the transformation. They are considered outliers. If we need a new feature point in this block (see the section about tracking) then point $\mathbf{u}^1$ from the pixel match with the maximum ZNCC among inlier correspondences within is selected as the representative point of the block and transferred to the second image with subpixel precision using estimated H. This is different from [Lhuillier and Quan 2005] where block centers are selected as representative points. Our approach is superior since block center may not correspond to a point with any image information or may be within the outlier subregion that is not fitted by homography.

Algorithm 1 Two-frame quasi-dense pixel match propagation

Input : Seed pixel matches
Output : Quasi-dense pixel matches in **Map**

Seeds - collection of matches to be propagated
Map - collection of matches sorted by ZNCC
LocalMap - collection of local matches sorted by ZNCC

Add seed matches to **Map** and **Seeds**

while **Seeds** is not empty {
  Pull match $(\mathbf{x}^1,\mathbf{x}^2)$ with maximum ZNCC from **Seeds**

  Clear **LocalMap**
  for all matches $(\mathbf{u}^1,\mathbf{u}^2)$ from $\mathscr{N}(\mathbf{x}^1,\mathbf{x}^2)$ {
    if $s(\mathbf{u}^1) > t$ and $s(\mathbf{u}^2) > t$ and
    $ZNCC(\mathbf{u}^1,\mathbf{u}^2) > z$ and
    $(\mathbf{u}^1,\mathbf{u}^2)$ fit the motion model
      Add $(\mathbf{u}^1,\mathbf{u}^2)$ to **LocalMap**
  }

  while **LocalMap** is not empty {
    Pull match $(\mathbf{u}^1,\mathbf{u}^2)$ with maximum ZNCC from **LocalMap**
    if neither $(\mathbf{u}^1,*)$ nor $(*,\mathbf{u}^2)$ are present in **Map**
      Add $(\mathbf{u}^1,\mathbf{u}^2)$ to **Map** and **Seeds**
  }
}

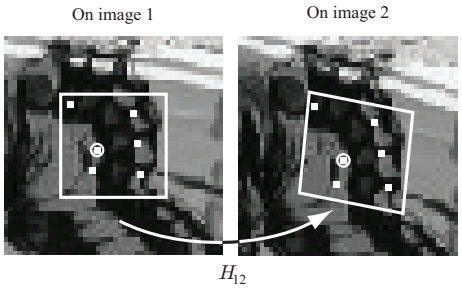

Figure 4: Block in the first image transferred to the second. Some pixel matches are shown (not all of them). Representative point is rounded by a circle in each image.

All these obtained quasi-dense subpixel correspondences are checked by fundamental matrix F or homography H that was robustly estimated from seed sparse points during the sparse algorithm stage (model is selected according to the GRIC [Konushin et al. 2005]). I.e. for pixel $\mathbf{u}^1$ corresponding pixel $\mathbf{u}^2$ must lie close to epipolar line $F\cdot\mathbf{u}^1$ or predicted position $H\cdot\mathbf{u}^1$ respectively:

$$dist_{lp}(F\cdot\mathbf{u}^1,\mathbf{u}^2) < dt_F \text{ or}$$
$$dist_{pp}(H\cdot\mathbf{u}^1,\mathbf{u}^2) < dt_H$$

$dist_{lp}$ denotes distance from line to point, $dist_{pp}$ denotes distance from point to point and dt_F, dt_H are corresponding thresholds (usually selected to be 1-2 pixels). GRIC selection of homography means either motion (camera rotates around a fixed point) or structure (all considered points are coplanar) degeneracy.

Quasi-dense 2 frame matching algorithm outline:
Input : Pair of images, seed pixel matches and F or H for this pair of frames.

1. Propagate seed matches to quasi-dense pixel matches using F or H for guided matching;
2. For each small block of the first image estimate a local affine homography that transfers quasi-dense pixel points within the block to their correspondences in the second image. Throw away pixel matches that do not fit this affinity;
3. Use estimated local affine homographies to transfer subpixel quasi-dense points within the block. These may be either newly detected representative point of the block or point tracked from the previous frame (see the section about tracking);
4. Subpixel point correspondences are checked against motion model and are thrown away if they do not fit.

Output : Local homographies and representative Quasi-dense point matches for each block

4.2 Quasi-dense matching for 3 frames

We propose quasi-dense matching for 3 frames that extends 2 frame approach and allows detection of inconsistent matches when epipolar constraint is too weak allowing erroneous displacement along the epipolar line.

Again Pixel-to-pixel-to-pixel (it's shortened to '3-pixel' hereafter) quasi-dense correspondences are obtained at first. They are established through recursive propagation of existing 3-pixel matches. This process is initialized by seed matches that are 3-frame sparse matches downsampled to pixel precision.

Quality of match $(\mathbf{u}^1,\mathbf{u}^2,\mathbf{u}^3)$ is measured by product of ZNCCs (we call it ZNCC3):

$$ZNCC(\mathbf{u}_1,\mathbf{u}_2)\cdot ZNCC(\mathbf{u}_2,\mathbf{u}_3)\cdot ZNCC(\mathbf{u}_1,\mathbf{u}_3)$$

That means that all submatches $(\mathbf{u}^1,\mathbf{u}^2)$, $(\mathbf{u}^2,\mathbf{u}^3)$, $(\mathbf{u}^1,\mathbf{u}^3)$ must be consistent. The threshold z on this product is set again to 0.5.

Cross-consistency check also naturally extends to 3 images: for each $\mathbf{u}^i$ ZNCCs with $\mathbf{u}^j$ and $\mathbf{u}^k$ must be maximum among possible matches on jth and kth image respectively ($\{i,j,k\}$ are permutations of $\{1,2,3\}$). *2d disparity gradient limit* is also employed and trifocal tensor (preestimated from sparse points) is used to predict position in the third image when supplied with positions in the other two images. It is used here instead of reprojection due to performance reasons.

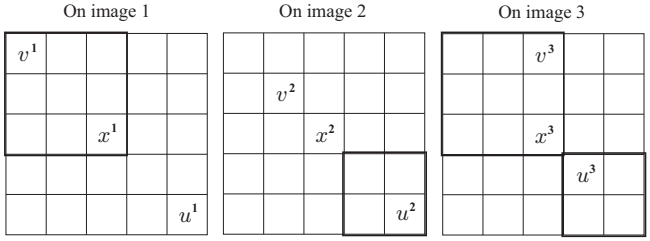

Figure 5: Neighborhood of $(\mathbf{x}^1,\mathbf{x}^2,\mathbf{x}^3)$ where possible matches for $\mathbf{u}^1$ are in the frames in the neighborhoods of $\mathbf{x}^2$ and $\mathbf{x}^3$,possible matches for $\mathbf{v}^2$ are in the frames in the neighborhoods of $\mathbf{x}^1$ and $\mathbf{x}^3$

Confidence measure is used in the same way as with two-frame matching.

Algorithm 2 Three-frame quasi-dense pixel match propagation

Input : Seed pixel matches
Output : Quasi-dense pixel matches in **Map**

Seeds - collection of matches to be propagated
Map - collection of matches sorted by ZNCC3
LocalMap - collection of local matches sorted by ZNCC3

Add seed matches to **Map** and **Seeds**

while **Seeds** is not empty {
  Pull match $(\mathbf{x}^1,\mathbf{x}^2,\mathbf{x}^3)$ with maximum ZNCC3 from **Seeds**

  Clear **LocalMap**
  for all matches $(\mathbf{u}^1,\mathbf{u}^2,\mathbf{u}^3)$ from $\mathcal{N}(\mathbf{x}^1,\mathbf{x}^2,\mathbf{x}^3)$ {
    if $s(\mathbf{u}^1) > t$, $s(\mathbf{u}^2) > t$, $s(\mathbf{u}^3) > t$ and
      $ZNCC3(\mathbf{u}^1,\mathbf{u}^2,\mathbf{u}^3) > z$ and
      $(\mathbf{u}^1,\mathbf{u}^2,\mathbf{u}^3)$ fits the motion model
        Add $(\mathbf{u}^1,\mathbf{u}^2,\mathbf{u}^3)$ to **LocalMap**
  }

  while **LocalMap** is not empty {
    Pull match $(\mathbf{u}^1,\mathbf{u}^2,\mathbf{u}^3)$ with max ZNCC3 from **LocalMap**
    if neither $(\mathbf{u}^1,*,*)$ nor $(*,\mathbf{u}^2,*)$ nor $(*,*,\mathbf{u}^3)$
    are present in **Map**
      Add $(\mathbf{u}^1,\mathbf{u}^2,\mathbf{u}^3)$ to **Map** and **Seeds**
  }
}

After quasi-dense pixel matches have been computed we obtain local homographies and resampled quasi-dense subpixel correspondences. We subdivide first image into rectangular blocks and robustly (using RANSAC) fit a local affine transformations H_{12} and H_{13} to pixel quasi-dense correspondences $(\mathbf{u}^1,\mathbf{u}^2)$ and $(\mathbf{u}^1,\mathbf{u}^3)$ respectively whose first point $\mathbf{u}^1$ is within the block assuming that most pixel matches within the block are approximately lie on the single planar patch [Lhuillier and Quan 2005]. RANSAC procedures provide information about pixel matches that do not fit to either of the transformations. They are considered outliers. If we need a new feature point in this block (see the section about tracking) then point $(\mathbf{u}^1$ from the pixel match with the maximum ZNCC3 among inlier correspondences within the block is selected as the representative point of this block and transferred to the second and the tird images with subpixel precision using estimated H_{12} and H_{13}.

The propagation algorithm outline is generally the same in the two-frame case, it is presented in Algorithm 2.

All these obtained quasi-dense subpixel correspondences are checked by trifocal tensor T or (what is equivalent and generally preferred) by reprojection of point triangulated from each pair among three frames to the third and measuring a distance to the tracked point position that must be less than a threshold (typically 1-2 pixels). If any of the reprojected point positions does not satisfy this constraint this feature is considered to be an outlier. I.e. for a match $(\mathbf{u}^1,\mathbf{u}^2,\mathbf{u}^3)$ the following constraints must be satisfied:

$$\begin{aligned} dist_{pp}(reproj\_from(\mathbf{u}^1,\mathbf{u}^2),\mathbf{u}^3) &< dt_T \text{ and} \\ dist_{pp}(reproj\_from(\mathbf{u}^2,\mathbf{u}^3),\mathbf{u}^1) &< dt_T \text{ and} \\ dist_{pp}(reproj\_from(\mathbf{u}^3,\mathbf{u}^1),\mathbf{u}^2) &< dt_T \end{aligned}$$

Threshold dt_T is usually selected to be 1-2 pixels.

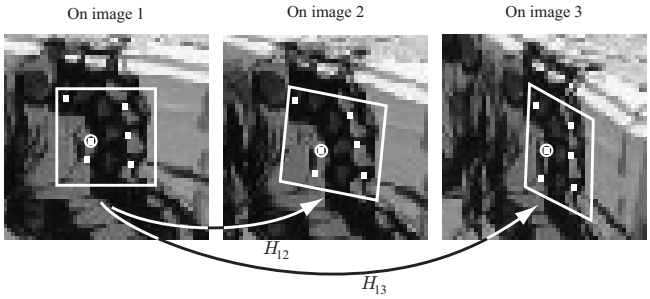

Figure 6: Block in the first image transferred to the second and the third. Some pixel matches are shown (not all of them). Representative point is rounded by a circle in each image.

Quasi-dense 3 frame matching algorithm outline:
Input : Triple of images, seed pixel matches and trifocal tensor T.

1. Propagate seed matches to quasi-dense pixel matches using T for guided matching;
2. For each small block of the first image estimate local affine homographies that transfers quasi-dense pixel points within the block to their correspondences in the second image and the third images. Throw away pixel matches that do not fit either of these affinities;
3. Use estimated local affine homographies to transfer subpixel quasi-dense points within the block. These may be either newly detected representative point of the block or point tracked from the previous frame (see the section about tracking);
4. Subpixel point correspondences are checked against motion model and are thrown away if they do not fit.

Output : Local homographies H_{12},H_{13} and representative Quasi-dense point matches for each block

4.3 Quasi-dense tracking for multiple frames

To achieve better reconstruction accuracy longer tracks are preferable. We establish quasi-dense tracks by means of the following procedure for each keyframe, which comes in two variants relying on two or three frame matching respectively that are described in the previous sections.

Subpixel quasi-dense pixel matches are obtained by two or three-frames algorithm for the last two or three frames including the current. That means that this algorithm starts with the first two or three keyframes from the sequence beginning and repeatedly executed for the consecutive pairs or triples of keyframes (see Figure 7). The last frame is denoted as 'current' hereafter.

In the two frame case we have a set of local affine transformations to the second image for each small block of the first image after quasi-dense pixel matching has completed. We transfer all quasi-dense feature points that are being tracked (i.e. were active on the previous frame) from the previous frame to the current using these local affine transformations. The quality of the correspondences is validated by checking against geometric constraints. In our stratified algorithm we use camera information obtained from the sparse stage and may predict point position by reprojecting 3d point triangulated from previous projections of this feature. But if this point was present on only two frames its triangulation accuracy may be poor.

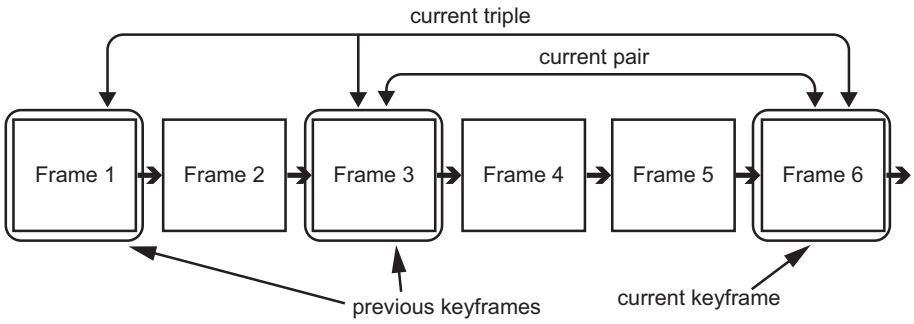

Figure 7: Current and previous frames for 2-frame and 3-frame based tracking

That's why we prefer an alternative approach when we consider last position as valid and perform triangulation for all keyframes including current. If the accuracy of reconstruction (measured by maximum error of reprojection for all frames where projection of this point is present) remains below an inlier threshold then point position in the current frame is considered approved. After all the existing points have been analyzed we look through blocks of the first frame (of this pair, not a sequence) where affine transformation was built but there were no existing points. We initialize new tracks in these blocks by selecting representative point in the block (see section 4.1) in the first frame and make a beginning of new track with this point and its position in the second image determined by local affinity transfer. This technique allows us to maintain density of quasi-dense correspondences through the sequence.

The three frame case is more complicated since pixel matches are established between three frames at once. There are possibly fewer pixel matches but they are more reliable. The other part is generally the same as for two frames except that we search for empty blocks in the first frame of the triple instead of a pair and newly added quasi-dense feature tracks will immediately contain three frames. The second and the third entries of the track are obtained using first-to-second and first-to-third local affine homographies.

5 Quasi-dense structure reconstruction

After quasi-dense tracking has been completed reconstruction of 3d quasi-dense points is obtained. As we already have all the cameras estimated on the sparse stage it is a straightforward multi-frame triangulation. Standard algorithm from [Hartley and Zisserman 2004] is employed for this task. After that bundle adjustment with point-only variation is performed.

6 Experimental results

We have studied behavior of the algorithms on a number of real world sequences captured by a hand-held camera (Canon IXUS 500). Quantitative evaluation results for two sample sequences (15 and 30 frames long respectively) are presented in the following tables:

Method	Number of points	RMS	Mean track length
SPARSE	1463	0.42	6.9
QUASI-LQ	6723	0.5	3.5
QUASI-2F	8213	0.46	3.8
QUASI-3F	7562	0.45	4.1

Method	Number of points	RMS	Mean track length
SPARSE	2743	0.37	7.4
QUASI-LQ	12245	0.47	3.8
QUASI-2F	14303	0.45	4.2
QUASI-3F	13117	0.42	4.6

SPARSE denotes sparse points tracking method from [Konushin et al. 2005]. QUASI-LQ denotes quasi-dense matching method from [Lhuillier and Quan 2002b]. QUASI-2F denotes proposed quasi-dense tracking with 2-frame matching method. QUASI-3F denotes proposed quasi-dense tracking with 3-frame matching method.

RMS denotes mean reprojection distance for all points for all cameras after euclidean bundle adjustment.

Sample frame and quasi-dense structure reconstruction are presented in Figure 8.

As can be clearly seen from the tables both QUASI-2F and QUASI-3F provide longer tracks than QUASI-LQ method due to more careful selection of initial quasi-dense features. Besides this QUASI-3F is superior to QUASI-LQ and QUASI-2F in terms of accuracy measured as RMS.

7 Conclusion and future work

In this paper a new guided quasi-dense structure estimation framework has been proposed. It has been demonstrated that it provides higher precision of 3d points estimations than that of other quasi-dense matching methods.

Our method differs from existing methods in several ways. First, we apply quasi-dense matching only to pairs and triples of adaptively selected key-frames, for which multi-view relation can be reliably estimated. Second, we use sparse feature tracks that uniformly distributed in images as seed matches for quasi-dense correspondence propagation. Third, we estimate camera movement from scene structure and use projective or Euclidian multi-view geometry to guide quasi-dense matching for 3-frames segments. Outliers in quasi-dense matches are segmented by thresholding reprojection error of corresponding 3d points.

Figure 8: Sample frame from a sequence and quasi-dense 3d reconstruction

About the authors

Andrei Khropov is PhD student in Moscow State University. He has received his specialist degree in mathematics in 2005 in MSU.

Anton Konouchine, PhD, has received his specialist degree in computer science in 2002 in MSU. He received his PhD in 2005 from Keldysh Institute of Applied Mathematics. He is now a research fellow in Moscow State University.

References

FISCHLER, M. A., AND BOLLES, R. C. 1981. Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. *Commun. ACM 24*, 6, 381–395.

FITZGIBBON, A. W., AND ZISSERMAN, A. 1998. Automatic camera recovery for closed or open image sequences. In *ECCV (1)*, 311–326.

GIBSON, S., COOK, J., HOWARD, T., HUBBOLD, R. J., AND ORAM, D. 2002. Accurate camera calibration for off-line, video-based augmented reality. In *ISMAR*, 37–46.

HARTLEY, R. I., AND ZISSERMAN, A. 2004. *Multiple View Geometry in Computer Vision*, second ed. Cambridge University Press, ISBN: 0521540518.

KONUSHIN, A., GAGANOV, V., AND VEZHNEVETS, V. 2005. Combined guided tracking and matching with adaptive track initialization. In *Graphicon*, 311–326.

LHUILLIER, M., AND QUAN, L. 2002. Match propogation for image-based modeling and rendering. *IEEE Trans. Pattern Anal. Mach. Intell. 24*, 8, 1140–1146.

LHUILLIER, M., AND QUAN, L. 2002. Quasi-dense reconstruction from image sequence. In *ECCV (2)*, 125–139.

LHUILLIER, M., AND QUAN, L. 2003. Surface reconstruction by integrating 3d and 2d data of multiple views. In *ICCV*, 1313–1320.

LHUILLIER, M., AND QUAN, L. 2005. A quasi-dense approach to surface reconstruction from uncalibrated images. *IEEE Trans. Pattern Anal. Mach. Intell. 27*, 3, 418–433.

POLLEFEYS, M., AND GOOL, L. J. V. 2002. From images to 3d models. *Commun. ACM 45*, 7, 50–55.

POLLEFEYS, M., KOCH, R., AND GOOL, L. J. V. 1998. Self-calibration and metric reconstruction in spite of varying and unknown internal camera parameters. In *ICCV*, 90–95.

POLLEFEYS, M., KOCH, R., VERGAUWEN, M., AND GOOL, L. J. V. 1999. Hand-held acquisition of 3d models with a video camera. In *3DIM*, 14–23.

QUAN, L. 1995. Invariants of six points and projective reconstruction from three uncalibrated images. *IEEE Transactions on Pattern Analysis and Machine Intelligence 17*, 1, 34–46.

SCHARSTEIN, D., AND SZELISKI, R. 2002. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. In *International Journal of Computer Vision*, vol. 47, 7–42.

THORMÄHLEN, T., BROSZIO, H., AND WEISSENFELD, A. 2004. Keyframe selection for camera motion and structure estimation from multiple views. In *ECCV (1)*, 523–535.

TOMASI, C., AND KANADE, T. 1991. Detection and tracking of point features. Tech. Rep. CMU-CS-91-132, Carnegie Mellon University, April.

TORR, P., AND MURRAY, D., 1997. The development and comparison of robust methods for estimating the fundamental matrix.

TORR, P., AND ZISSERMAN, A., 1997. Robust parameterization and computation of the trifocal tensor.

TORR, P. H. S., AND ZISSERMAN, A. 1999. Feature based methods for structure and motion estimation. In *Workshop on Vision Algorithms*, 278–294.

TRIGGS, B., MCLAUCHLAN, P. F., HARTLEY, R. I., AND FITZGIBBON, A. W. 1999. Bundle adjustment - a modern synthesis. In *Workshop on Vision Algorithms*, 298–372.

ZENG, G., PARIS, S. X., QUAN, L., AND LHUILLIER, M. 2004. Surface reconstruction by propagating 3d stereo data in multiple 2d images. In *ECCV (1)*, 163–174.