Pacific Graphics 2014
J. Keyser, Y. J. Kim, and P. Wonka
(Guest Editors)

*Volume 33 (2014), Number 7*

# Making in-Front-of Cars Transparent: Sharing First-Person-Views via Dashcam

Shao-Chi Chen[1] Hsin-Yi Chen[1] Yi-Ling Chen[1,2] Hsin-Mu Tsai[1,2] Bing-Yu Chen[1,2]

[1]National Taiwan University [2]Intel-NTU Connected Context Computing Center

(a) (b) (c)

**Figure 1:** *A synthesis result of the proposed view sharing system. (a) The original view of the preceding vehicle. (b) The original view of the subject vehicle with a large portion of the image blocked by the preceding vehicle. (c) The perspective of the preceding vehicle is transferred to the corresponding view of the subject vehicle to "disocclude" the blocked area as if the preceding vehicle becomes transparent.*

**Abstract**

*Visual obstruction caused by a preceding vehicle is one of the key factors threatening driving safety. One possible solution is to share the first-person-view of the preceding vehicle to unveil the blocked field-of-view of the following vehicle. However, the geometric inconsistency caused by the camera-eye discrepancy renders view sharing between different cars a very challenging task. In this paper, we present a first-person-perspective image rendering algorithm to solve this problem. Firstly, we contour unobstructed view as the transferred region, then by iteratively estimating local homography transformations and performing perspective-adaptive warping using the estimated transformations, we are able to locally adjust the shape of the unobstructed view so that its perspective and boundary could be matched to that of the occluded region. Thus, the composited view is seamless in both the perceived perspective and photometric appearance, creating an impression as if the preceding vehicle is transparent. Our system improves the driver's visibility and thus relieves the burden on the driver, which in turn increases comfort. We demonstrate the usability and stability of our system by performing its evaluation with several challenging data sets collected from real-world driving scenarios.*

Categories and Subject Descriptors (according to ACM CCS): I.4.9 [Image Processing and Computer Vision]: Applications—

## 1. Introduction

Motivated by advent of cost effective and widely available camcorders, nowadays it is common to see a car driver using a dashcam (dashboard camera), a portable camera that is attached to the interior of the windshield, to record videos capturing objects in front of the car when in motion. In the unfortunate event that the car is involved in an accident, the recorded videos can serve as evidence for insurance and le-

**Figure 2:** *Previously proposed See-Through System (STS) presents to the driver a view with images taken from the preceding vehicle directly super-imposing over the image area occupied by the preceding vehicle [GVF12]. However, the drivers need to pay extra attention since the perceived contextual information from different views are highly inconsistent.*

gal purposes. Since the dashcam can be treated as a type of first-person-view of the car, instead of using it only as a passive record, in this paper we develop a solution to utilize other vehicles' views (i.e., taken from their dashcams) to improve driver perception and increase the level of driving safety.

Considering a vision-obstructing large vehicle in front of ours while driving, critical decisions such as lane changing or overtaking cannot be easily made because drivers cannot be fully aware of the potential dangers behind the visual obstruction. Although it has been shown that the overtaking vehicle can utilize direct vehicle-to-vehicle (V2V) communications to access the video data recorded by the front vehicle without significant delay [GVF12], rendering the video streaming in the perspective of preceding vehicle requires the overtaking drivers to continuously pay attention to two disjointed views in different perspectives. Such fragmented views and inconsistent perspectives cause degradation in spatial cognition and place extra burden on the overtaking driver. For example, in Figure 2, the visual discontinuities around the boundaries of the darkened rear windscreen are not only distracting but could also have a negative impact on driving safety.

Given two synchronized video sequences $I^r$ and $I^t$, which are captured by a preceding vehicle ($r$) and the subject vehicle ($t$), respectively. The field-of-view in $I^t$ are partially obstructed by the preceding vehicle. Our goal is thus to generate an image sequence $\hat{I}^t$, where the occluded regions in each frame of $I^t$ are replaced by the visible visual elements appearing in the corresponding frame of $I^r$ with the perspective of the subject vehicle. To produce $\hat{I}^t$, a straightforward solution is to perform a pairwise image matching and stitching between two corresponding frames, as suggested by [BL07]. However, the performance of such process is affected by the following difficulties. Firstly, if the subject vehicle followed the preceding vehicle with a short distance, the occluded regions severely downgrade the matching quality. Secondly, the inconsistent parallax from scene depth and different camera locations violate the assumptions made in typical stitching approaches [SS97, BL07]. Finally but not lastly, applying the methods designed for images to process videos may lead to temporal artifacts, e.g., ghost effects of different misaligned objects in the video.

To address the above limitations and challenges, we propose a view-sharing system to integrate spatial information across two temporally aligned sequences. The proposed system performs both shape adjustment and color blending to generate the composited video such that the viewing perspectives and color appearances among different views are seamlessly fused and the temporal coherence can be also achieved. To this end, we propose a video-based perspective adaptation technique consisting of two main steps: *local homography estimation* and *perspective-aware warping*. With our approach, the unobstructed view and perspective of the preceding vehicle can be gradually transformed and adapted to the matched occluded region in the subject video. Specifically, our approach makes use of the coherence of scene dynamics to guide the local warping across the long video sequences. We also allow local homographies to be accumulated to accelerate incremental homography propagation. In addition, the parallax problem is also handled properly by restricting image stitching within a local region.

In summary, the contributions of this work are stated as follows. Firstly, we propose a view-sharing system that integrates spatial information across two temporally synchronized dashcams. The generated video sequence enables the subject driver to monitor surroundings ahead of the obstructed vehicle in accordance with current visual perception, thus providing complete situational awareness that facilitates decision making and responses to driving events. Secondly, we exploit scene dynamics in a video and propose a spatially varying warping technique for locally adapting the visibility as well as the perspective of the preceding vehicle to the occluded region in the target location. It allows the subject driver to exceed the limited spatial visibility in a perspective-consistent way. Finally, we show that our system is of high practical value by evaluating it in different scenarios, including straight-lane regions on highways and curved-lane regions in urban areas.

## 2. Related Work

**The See-Through System.** Utilizing direct V2V communications and camera sensors in modern vehicles, Ferreira *et al.* [GVF12] developed the See-Through System (STS) to mitigate the driving difficulty in perceiving incoming traffic behind a vision-obstructing vehicle. The system recognizes the back of the preceding vehicle, and replaces it with a video feed from the dashcam mounted on the preceding vehicle. Despite the fact that V2V communications ensures minimal delay in obtaining the images of the preceding ve-

hicle [OMGF*10], the perspectives of the reproduced video were not adjusted in accordance with current visual perception. In this work, we propose a video-based perspective warping technique to adjust the perspective structure recorded by the preceding vehicle to adapt to the perspective of the temporally corresponding frame captured by the subject vehicle.

**Image stitching.** The goal of image stitching is to seamlessly integrate multiple images into a single mosaic. Conventional methods [SS97, BL07] assume that the images to be stitched together are associated with a global homography transformation, and thus work well with specific types of images, such as those of distant scenes or those taken from a camera rotated about its center of projection. However, global homograhy alone is not flexible enough to model all types of scenes and camera motions. Therefore, Gao *et al.* [GKB11] proposed a dual-homography model to address the insufficiency of global homography model. Still, this model cannot fully represent the diverse scene variations of real-world images. Recently, [LLM*11, ZCBS13] have proposed to use spatially-varying warping functions to account for parallax. Some other related work [AZP*05, DPR05] aimed to stitch images into panoramas from a video sequence. Although these methods have achieved great performance for stitching overlapping and consecutive images from a rotating camera, mosaicking views from translated cameras along a vehicle path remains a very challenging task due to the motion parallax in scene depth.

**Motion estimation.** Motion estimation techniques in video can be roughly classified into two categories: direct methods and feature-based methods. The former [LK81] produces dense correspondence, but it is not robust when objects present in one image but not in the other. The latter, such as [TCGP09], extracts visual features and tracks them across multiple frames. The benefit of this category is that it provides robust tracking when there is significant image motion. In addition, motion estimation by sparse feature tracking is useful for a variety of applications, i.e., video stabilization [LCCO09], video resizing [WFS*09] or video inpainting [GKT*12]. Different from their methods, we do not assume a restricted camera model as in video resizing or inpainting. Instead, we aim at perspective adaptation, where feature trajectories are exploited to compute spatially varying warping functions to guide local warping.

**Video alignment.** The primary goal of sequence-to-sequence alignment is to establish both spatial and temporal correspondences of the same dynamic scene. Although this problem has been extensively studied in the literature [CI02, ST04, EB13], they assume that there is sufficiently large overlap between spatially corresponding frames. Under such assumption, the occlusion problem is thus minimized or can be ignored. In this case, video alignment is only suitable for very similar scenes. In our setting, we aim to register two related frames with substantially different appearance due to occlusion or change of perspective, presenting new challenges to current video alignment approaches.

## 3. Overview

Figure 3 depicts the overall algorithmic flow of our video perspective warping technique. The input to our system consists of a *target* sequence and a *reference* sequence, which are assumed to be temporally synchronized. The system first estimates the vision-obstruction regions (Figure 3(a)) in the target sequence (Section 4.1). To generate the corresponding contour of the visible visual region captured by the reference image, we track the robust feature trajectories through the spatial-temporal volume in the reference sequence, as described in Section 4.2. The area inside the contour (Figure 3(b)) is then transformed across multiple frames by the proposed perspective adaptation algorithm (Figure 3(c)) and stitched to the matched occluded region in the target frame (Figure 3(d)). To avoid perceptional discrepancy and mismatched boundaries between the transformed region and the target image, our perspective adaptation algorithm (Section 4.3) adjusts the shape of the transformed region so that the viewpoints are continuous across the boundaries of the transformed region and the target image. Specifically, spatially-varying warping and local stitching process are performed in the area inside the contour through the video volume until the transformed region adapts to the viewpoint of the target image. Figure 3(e) shows the final synthesized image which is seamlessly composited from the reference and the target images and achieves consistent visual appearance along the boundaries and perspective projection.

### 3.1. Problem Formulation

Considering two moving vehicles, namely the subject vehicle and the preceding vehicle. Denote the *target* and the *reference* sequences captured by the subject vehicle and preceding vehicle as $I^t(\mathbf{x})$ and $I^r(\hat{\mathbf{x}})$, respectively. For each frame in $I^t(\mathbf{x})$, the captured scene is partially occluded by the preceding vehicle. Let $\mathbf{x} = [x, y, m]$ and $\hat{\mathbf{x}} = [\hat{x}, \hat{y}, n]$ denote the spatial-temporal coordinates of $I^t$, $I^r$ with $m = 1, ..., M$ and $n = 1, ..., N$ indicating their frame indices, respectively. For simplicity, we will refer to the $m$-th target and $n$-th reference frame as $I^t_m$ and $I^r_n$ in the following discussions. Furthermore, we assume that the temporal mapping is expressed through a discrete-time signal mapping function $T : N \rightarrow R$, such that $(m, T(m))$ is an assignment of an target frame to a reference frame. For each input frame $I^t_m$, the temporal mapping $(m, T(m))$ is assumed to be determinable in real-time via a wireless vehicular communication system. For each frame $I^t_m$ in the target sequence, our goal is to replace the vision-obstruction region with the visual elements in the temporally corresponding frame $I^r_{T(m)}$ according to the perspective projection of $I^t_m$. Specifically, it will create an im-

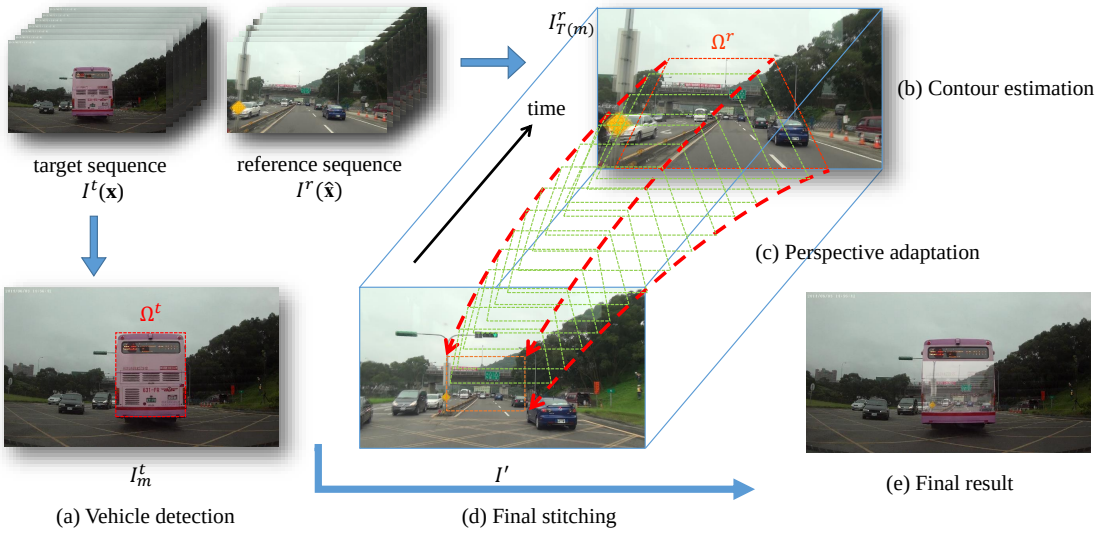



**Figure 3:** *An overview of the proposed method. Given the target and reference sequences, the occluded region (a) in the target image is estimated and our system automatically finds the corresponding contour (b) in the reference image. To transform the area inside the contour in the reference image to match the occluded region in the target image, the perspective of the region to be transformed are adapted to fit that of the location in the target image by performing (c) perspective adaptation through reference video volume and (d) a stitching process between the two image frames. In the stage of perspective adaptation, a novel view $I'$ is synthesized by performing local homography estimation and perspective-aware warping. Finally, we stitch the synthesized view and target image where the warped region is seamlessly blended into the target image to make an impression that the vehicle is transparent (e). Note that the "see-through" effect does not cover the entire occluded region such that the viewers remain consciously aware of the existence of the preceding vehicle, thus improving driving safety.*

pression as if the preceding vehicle becomes transparent, as shown in Fig. 1.

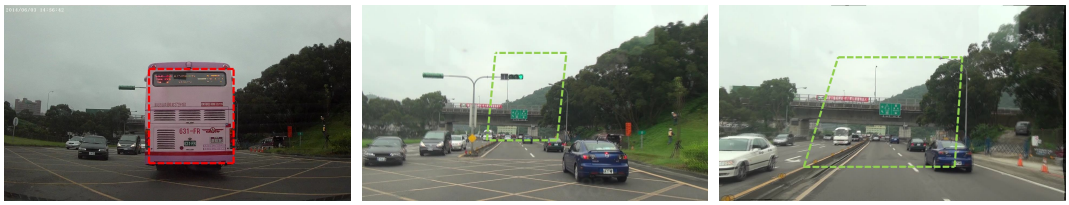
## 4. Method

### 4.1. Occlusion detection

In the target sequence, the occluded region corresponds to the preceding vehicle's positions. To contour such region $\Omega^t_m$ in each frame $I^t_m$, the robust object tracking method proposed by Zoung *et al.* [ZLY12] is adopted to obtain an accurate vehicle position. The tracker is initialized by a vehicle object detector, then the states of the target position is estimated and updated using a collaborative model. The method outperforms many other object tracking methods [WLY13] when the scale variation is large, e.g., the preceding vehicle has sudden movements or the relative speed between the two vehicles changes irregularly, which are both of common situations when cars are in motion.

### 4.2. Contour generation

After the occluded position $\Omega^t_m$ (Figure 4(a)) in the target image $I^t_m$ is estimated, the goal of this stage is to generate the corresponding contour $\Omega^r_{T(m)}$ in the reference image $I^r_{T(m)}$ that contains the visual elements invisible in the target image. Directly matching $I^r_{T(m)}$ and $I^t_m$ is infeasible since $\Omega^t_m$ is not visible in $I^r_{T(m)}$. Therefore, we propose to perform forward tracking in the reference video volume to locate the position of such a region. For each target frame $I^t_m$, GPS information is firstly utilized to find the spatially closest frame $I^r_{T(m)-k}$ (Figure 4(b)) in reference sequence, where $k$ is an integer index offset. GPS alignment guarantees that the corresponding inter-sequence frame pair $(I^t_m, I^r_{T(m)-k})$ is taken approximately at the same geographical locations within the range of about $2.5 \sim 5$ meters. Next, we perform the traditional image matching technique on the image pair $(I^t_m, I^r_{T(m)-k})$ to estimate a global transformation between the two images. Since they are captured from similar viewpoints, a global transformation model is sufficient to roughly model their transformation. Then, we use the global transformation to warp $\Omega^t_m$ to $I^r_{T(m)-k}$ as the initial contour $\Omega^r_{T(m)-k}$ for the following forward tracking process. To find the estimated contour $\Omega^r_{T(m)}$ in $I^r_{T(m)}$, starting from $I^r_{T(m)-k}$, we detect features within $\Omega^r_{T(m)-k}$ and recover their trajectories by making a forward sweep through the reference video volume. Figure 4(c) shows the result of the estimated contour $\Omega^r_{T(m)}$. As will be explained in Section 4.3, the visual elements in the estimated region $\Omega^r_{T(m)}$ is then gradually transformed and aligned with the input image.



(a) (b) (c)

**Figure 4:** *(a) Target image $I^t_m$ and the occluded region $\Omega^t_m$ (indicated by the red dotted line). (b) The spatially corresponding frame $I^r_{T(m)-k}$ of $I^t_m$ obtained by utilizing the GPS information. The corresponding position of $\Omega^t_m$ is estimated by a global transformation (green dotted line) (c) The generated contour $\Omega^r_{T(m)}$ in $I^r_{T(m)}$ by our method.*

### 4.3. Perspective adaptation

Given the target image $I^t_m$ with the occluded region $\Omega^t_m$ and the reference image $I^r_{T(m)}$ with the contour mask $\Omega^r_{T(m)}$, specifying the region to be transformed and to fill into the occluded region in $I^t_m$, the goal of perspective adaptation is to synthesize a novel view which adapts to both the shape and the perspective of the target image while closely approximating the original local appearance of the transformed region.

An important characteristic of perspective projection is *foreshortening*: objects become smaller as their distances from the observer increase [LSC*12]. In other words, the projected size of an object is highly dependent on its depth. Typically, for a moving camera, the depth of the captured scene changes gradually. An important observation is that the change of appearance between consecutive frames also reveals how the perspective changes. When there is a significant discrepancy between the perspectives of the target and the reference images, the 2D shape of the transformed region must be adjusted according to the adapted motion to match such changes or discrepancies. To this end, we propose the following perspective-adaptation technique to accomplish this task.

#### 4.3.1. Perspective-aware warping and stitching

Rendering a consistent perspective view can be achieved by estimating the transformation function between the reference and the target images. Specifically, given the estimated homography $\mathbf{H} \in \mathbb{R}^{3\times3}$, a pixel at position $\hat{\mathbf{x}} = [\hat{x}, \hat{y}]^T$ in the reference image $I^r_{T(m)}$ is warped to the position $\mathbf{x} = [x, y]^T$ in the target image $I^t_m$ by

$$\mathbf{x}' = \mathbf{H}\hat{\mathbf{x}}', \tag{1}$$

where $\mathbf{x}'$ is $\mathbf{x}$ in homogeneous coordinates. In inhomogeneous coordinates,

$$x = \frac{\mathbf{h}_1^T[\hat{x}\ \hat{y}\ 1]^T}{\mathbf{h}_3^T[\hat{x}\ \hat{y}\ 1]^T} \quad \text{and} \quad y = \frac{\mathbf{h}_2^T[\hat{x}\ \hat{y}\ 1]^T}{\mathbf{h}_3^T[\hat{x}\ \hat{y}\ 1]^T}, \tag{2}$$

where $\mathbf{h}_j^T$ is the $j$-th row of $\mathbf{H}$. Eq. 2 can be rewritten as:

$$\mathbf{0}_{3\times1} = \begin{bmatrix} \mathbf{0}_{1\times3} & -\hat{\mathbf{x}}'^T & y\hat{\mathbf{x}}'^T \\ \hat{\mathbf{x}}'^T & \mathbf{0}_{1\times3} & -x\hat{\mathbf{x}}'^T \\ -y\hat{\mathbf{x}}'^T & x\hat{\mathbf{x}}'^T & \mathbf{0}_{1\times3} \end{bmatrix} \mathbf{h}, \quad \mathbf{h} = \begin{bmatrix} \mathbf{h}_1 \\ \mathbf{h}_2 \\ \mathbf{h}_3 \end{bmatrix}. \tag{3}$$

Let $\mathbf{a}_i \in \mathbb{R}^{2\times9}$ be the first two rows of Eq. 3 computed for the $i$-th correspondence pair $\{\mathbf{x_i}, \hat{\mathbf{x}}_i\}$. Direct Linear Transformation (DLT) is one of the techniques to estimate the nine elements of $\mathbf{H}$ from a set of correspondences $\{\mathbf{x}_i, \hat{\mathbf{x}}_i\}_{i=1}^N$ by

$$\mathbf{h} = \arg\min_{\mathbf{h}} \sum_{i=1}^{N} \|\mathbf{a}_i\mathbf{h}\|^2 = \arg\min_{\mathbf{h}} \|\mathbf{A}\mathbf{h}\|^2, \tag{4}$$

where $\mathbf{A} \in \mathbb{R}^{2N\times9}$ is obtained by stacking vertically $\mathbf{a}_i$ for all $i$. The solution is the least significant right singular vector of $\mathbf{A}$. Although a single 2D global transformation performs well for planar scenes or rotational camera motions, but for complex scenes, i.e., highly non-planar scene that is captured by different cameras in vehicle paths, as in our situation, the assumptions on motion properties and selection of dominant motions often lead to inaccurate results (Figure 5(c)). Moreover, due to the presence of occlusion, accurately aligning the input image and the reference image is much more challenging.

To tackle this obstacle, we propose a two-stage perspective-aware warping technique that utilizes the coherence of video dynamics to guide the perspective adaptation. In the first stage, we estimate the spatially varying warping functions between the consecutive frames in the reference sequence that describe how the transformed region (the area inside the contour $\Omega^r_{T(m)}$) should be deformed so that its size and shape matches the perspective of the target image. A novel view $I'_{T(m)-k}$ that integrates the visual element of transformed region while approximating the viewpoint of target image is synthesized by proceeding the process consecutively until the transformed region is gradually warped to $I^r_{T(m)-k}$, which represents the spatially closest frame of the target frame $I^t_m$ in the reference sequence. In the second stage, we align the synthesized image $I'_{T(m)-k}$

and the target image $I_m^t$ together, then the final composited image is recovered by blending the elements in the aligned image and the target image in the occluded region.

**Feature tracking.** The transformation model that relates the two images is typically estimated from noisy correspondences of local invariant features. Since consecutive video frames are usually very similar, we adopt the sparse optical flow method [ST94] to match corresponding feature points between two neighboring frames. The sparse optical flow method estimates the motion with a selected number of pixels,and thus it provides more robustness against noise than that of optical flow algorithms while avoiding the high computational cost due to frame-to-frame matching by using robust feature descriptors, i.e., SIFT [Low04]. Specifically, we compute interest points (Shi-Tomasi features) in the video frame and generate matched points for these interest points by tracking them across multiple frames. The tracking process produces fairly accurate matching results.

**Spatial varying warping function.** Let $\{\mathbf{x}_i, \tilde{\mathbf{x}}_i\}_{i=1}^N$ be the collected correspondence set across consecutive frames $I_t^r$ and $I_{t-1}^r$ in the reference sequence, where $\mathbf{x} = [x, y]$, $\tilde{\mathbf{x}} = [\tilde{x}, \tilde{y}]$, and $N$ is the number of correspondence pairs. To align two frames, a pixel at position $\mathbf{x}_*$ in frame $I_t^r$ is warped to the position $\tilde{\mathbf{x}}_*$ in frame $I_{t-1}^r$ by a location dependent homography model [ZCBS13]:

$$\tilde{\mathbf{x}}_* = \mathbf{H}_* \mathbf{x}_*, \tag{5}$$

where $\mathbf{H}_*$ is estimated from a weighted minimization problem:

$$\mathbf{h}_* = \arg\min_{\mathbf{h}} \| \sum_{i=1}^{N} \omega_*^i \mathbf{a}_i \mathbf{h} \|^2, \tag{6}$$

subject to $\|\mathbf{h}\| = 1$ and the weights $\{\omega_*^i\}_{i=1}^N$ are calculated from a Gaussian-like distribution:

$$\omega_*^i = \exp(-\frac{\|\mathbf{x}_* - \mathbf{x}_\mathbf{i}\|^2}{\sigma^2}), \tag{7}$$

where $\sigma^2$ is the variance. Eq. 7 gives higher weight to data points closer to $\mathbf{x}_*$. The problem can be written in the matrix form:

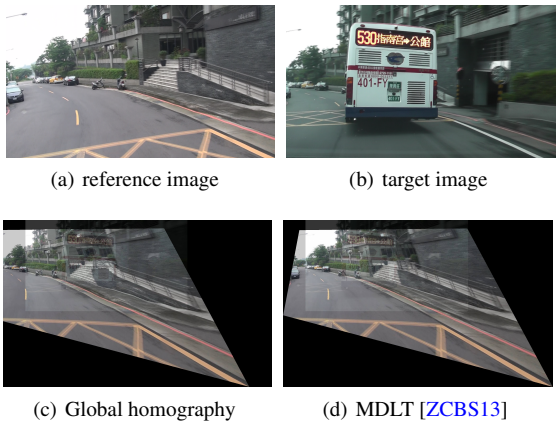
$$\mathbf{h}_* = \arg\min_{\mathbf{h}} \|\mathbf{W}_* \mathbf{A} \mathbf{h}\|^2, \tag{8}$$

where the $\mathbf{W}_* \in \mathbb{R}^{2N \times 2N}$ can be further described as:

$$\mathbf{W}_* = diag([\omega_*^1\ \omega_*^1 \ldots \omega_*^N \omega_*^N]). \tag{9}$$

$diag()$ constructs a diagonal matrix with a given vector. Eq. 8 corresponds to a weighted Singular Value Decomposition (WSVD) problem, and the solution is the least significant right singular vector of $\mathbf{W}_* \mathbf{A}$.

**Avoiding parallax using local stitch.** As mentioned in [ZL14], the images with significant parallax often cannot



(a) reference image (b) target image

(c) Global homography (d) MDLT [ZCBS13]

**Figure 5:** *Aligned images. (a) reference image captured by the preceding vehicle. (b) target image captured by the subject vehicle. (a) and (b) are input pairs. (c) the synthesized result stitched with global homography after final stitching. (d) the synthesized result stitched with MDLT method after final stitching.*

be aligned well over the whole overlapping region without suffering artifects like folding-over. To handle parallax, we also perform local stitch between $I_t^r$ and $I_{t-1}^r$. Specifically, after the local homographies between $I_t^r$ and $I_{t-1}^r$ are estimated, only the area inside the transformed contour $\Omega_t^r$ is warped to $I_{t-1}^r$, then a novel view $I_{t-1}^{\prime}$ is composited, which corresponds to the perspective observed in $I_{t-1}^r$. By iteratively applying local homography estimation and perspective-aware warping between $I_t^{\prime}$ and $I_{t-1}^r$, the content and the perspective inside the contour $\Omega_{T(m)}^r$ is gradually adjusted. Finally, a novel frame $I_{T(m)-k}^{\prime}$ is synthesized.

#### 4.3.2. Final stitching

Owing to the first warping stage discussed in Section 4.3.1, the perspective of $I_{T(m)-k}^{\prime}$ is adapted to $I_{T(m)-k}^r$, which is the spatially closest frame of the subject frame $I_m^t$ in reference video sequence. We assume that the perspectives of two frames should be similar if the distance between their spatial coordinates is small enough. Finally, we stitch $I_m^t$ and $I_{T(m)-k}^{\prime}$ together to get the final panoramic image $\hat{I}$. In order to make the preceding vehicle transparent, we only cut the part that corresponds to the occlusion mask $\Omega^t$ from the stitched view and blend it with $I_m^t$ to get the final result $\hat{I}$. We conclude this section by summarizing our approach in Algorithm 1.

## 5. Results and Discussions

We evaluate the performance of our system using video clips collected in real driving scenarios. The videos were designed to be captured in three different road conditions and traffic

**Algorithm 1** Algorithm

1: **Input:** target sequence $I^t$ and reference sequence $I^r$
2: **for** each $I^t_m, m = 1 \sim M$ **do**
3: $\Omega^t_m \leftarrow$ find the occluded region in $I^t_m$ (Sec. 4.1)
4: $I^r_{T(m)-k} \leftarrow$ find the spatially-closest frame in reference sequence with GPS information
5: $\Omega^r_{T(m)} \leftarrow$ estimate the corresponding contour in $I^t_{T(m)}$ (Sec. 4.2)
6: $I'_{T(m)-k} \leftarrow$ synthesize the novel view by perspective-aware warping and stitching (Sec. 4.3.1)
7: $\hat{I}_m \leftarrow$ obtain the final result by stitching $I'_{T(m)-k}$ and $I^t_m$ (Sec. 4.3.2)
8: **end for**
9: **Output:** synthesized sequence $\hat{I}$

flows: (i) on the highway, (ii) on the city road and (iii) on the mountain road. The dashcams on the vehicle were set up in the middle of the windshields with timestamps information. For capturing these videos, the driver on the subject vehicle followed in the path of the bus ahead, which corresponded to the lead vehicle through our discussion. In addition, the geographical information provided by GPS receivers on the vehicles was used to spatially align two videos. Beside, we also recorded an additional video with only one dashcam. In the following sections, we first demo the effectiveness of the proposed perspective adaptation approach in 5.1. Then, we compare our method with two approaches: (i) global alignment using RANSAC and (ii) local alignment with Moving Direct Linear Transformation and demonstrate several result using the proposed method in Section 5.2.

### 5.1. Verification of perspective adaptation

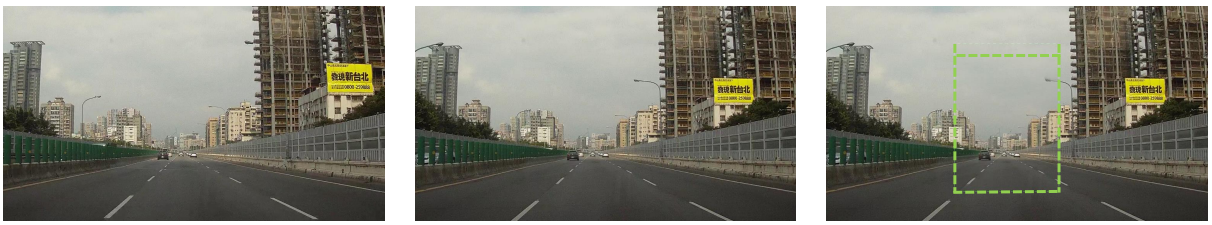
In this experiment, we aim to use a single video sequence to validate the effectiveness of the proposed perspective adaptation approach. In Figure 6(a), by gradually warping the area specified by the estimated contour at time $T(m)$ (obtained by forward propagation as described in Section 4.2), its perspective is adjusted to fit to that of time $T(m)-k$ and the image content can be seamlessly composited into the corresponding mask. The synthesis result is shown in Figure 6(c) and Figure 6(b) can be regarded as the ground truth (the frame captured at time $T(m)-k$). It can be seen that we alter the shape of transferred region according to scene depth change to model the perspective effect.

### 5.2. Qualitative comparisons

We compare our method against the baseline warping method (global homography via DLT in inliers) and the local homography method (Moving-DLT) [ZCBS13] with three alignment instances. Figure 7(a) and (b) show three pairs of input images with a significant amount of parallax and occlusion, where the input are the reference and target images captured by the dashcams on the lead vehicle and the subject vehicle, respectively. In each case, large viewpoint change, different lighting conditions and the presence of occlusion make the alignment task very challenging. Figure 7(c)∼(e) show the results generated by each method.

For the baseline method, we detect and match SIFT keypoints in the input pair, then run RANSAC to remove outliers. We estimate a global homography via Direct Linear Transformation (DLT) on inliers to align two images. The result of baseline method caused unavoidable misalignment and ghost effect, which can been seen in Figure 7(c). It suggests that using a single homography alone is not sufficient to model the transformations between two images because a scene is usually composed by more than two projection planes. Besides, given input images with considerably different perspectives, it's very difficult to establish enough correct matches thus leading to incorrect homography estimation. While Moving-DLT with spatially varying homographies is able to produce good results, it tries to align two images over the whole overlapping region. Thereby, the estimated transformations are easily dominated by the noisy matches. As a consequence, the distortion is very large and ghosting still occurs in the region Figure 7(d) (the stairs next to the wall). In contrast, by using the coherence property in the video sequence, features are easily to be matched, it facilitate the a good in first warping stage. In the stage of second warping stage, the perspective-adapting frame which possesses similar viewpoint of the subject vehicle is transformed. Thereby, it's easier to find matches between close perspective compared with direct warping methods. Therefore, our method can estimate local homography more precisely, thus achieves more plausible results as shown in Figure 7(e).

**Limitations** One limitation of the proposed method is that we have not considered inter-frame motion. Motion parallax must be carefully dealt with since it will cause the estimation of local homographies to diverge and compromise the spatial smoothness of the overall warping results. For future work, we would like to identify the object and camera motions since the MDLT and parallax-tolerant methods were designed for static scenes. To achieve this, we will investigate the use of advanced vehicle detection methods and other prior knowledge, e.g. the relative speed between different vehicles, to separate the moving objects from the static scene. The performance of our work is also highly dependent on the quality of feature detection and tracking. Its performance degrades in large textureless regions, such as ground and sky, since there are too few reliable feature points available to guide image warping and stitching. To overcome this limitation, we consider to detect planar surfaces in the scene such that a single homography can be more robustly estimated to guide image warping in these challenging scenarios.



(a) (b) (c)

**Figure 6:** *(a) is the frame $I^r_{T(m)}$ in the reference sequence. (b) is the frame $I^r_{T(m)-k}$ in the reference sequence. (In this case, we set k = 30). (c) is a novel image with the region inside a mask (green dotted line) synthesized by the proposed perspective adaptation method. One can see that the perspectives of the two different frames are very visually similar inside the mask.*

## 6. Conclusion

In this paper, we consider the problem of aligning two videos that are captured simultaneously by independently moving cameras following similar trajectories. Aligning two temporal synchronized video sequences encounters great challenges due to large viewpoint changes and heavy occlusion. Therefore, in this paper we propose a two-stage warping technique that gradually adapts the perspective from one video to the other, rather than directly aligning two videos with large difference in viewpoint. It not only reduces the difficulties of perspective transferring between multiple views, but also increases the visibility of the driver and enhances safety and comfort in driving scenarios.

## Acknowledgement

This work was partly supported by Ministry of Science and Technology, National Taiwan University and Intel Corporation under Grants MOST102-2911-I-002-001 and NTU103R7501.

## References

[AZP*05] AGARWALAA A., ZHENG C., PAL C., AGRAWALA M., COHEN M., CURLESS B., SALESIN D., SZELISK R.: Panoramic video textures. *ACM Trans. Graphics 24*, 3 (2005), 821–827. 3

[BL07] BROWN M., LOWE D. G.: Automatic panoramaic image stitching using invariant features. *ACM Trans. Graphics 74*, 1 (2007), 59–73. 2, 3

[CI02] CASPI Y., IRANI M.: Spatio-temporal alignment of sequences. *IEEE Transactions on Pattern Analysis and Machine Intelligence 24*, 11 (2002), 1409–1424. 3

[DPR05] DREW S., PAL C., RICHARD S.: Efficiently registering video into panoramic mosaics. In *Proc. IEEE Intl. Conf. Computer Vision* (2005), pp. 1300–1307. 3

[EB13] EVANGELIDIS G., BAUCKHAGE C.: Efficient subframe video alignment using short descriptors. *IEEE Trans. Pattern Analysis and Machine Intelligence 35*, 10 (2013), 2371–2386. 3

[GKB11] GAO J., KIM S. J., BROWN M. S.: Constructing image panoramas using dual-homography warping. In *Proc. IEEE Intl. Conf. Computer Vision and Pattern Recognition* (2011), pp. 49–56. 3

[GKT*12] GRANADOS M., KIM K. I., TOMPKIN J., KAUTZ J., THEOBALT C.: Background inpainting for videos with dynamic objects and a free-moving camera. In *Proc. European Conf. Computer Vision* (2012), pp. 682–659. 3

[GVF12] GOMES P. E. R., VIEIRA F., FERREIRA M.: The see-through system: From implementation to test-drive. In *Proc. IEEE Vehicular Networking Conf.* (2012), pp. 40–47. 2

[LCCO09] LEE K.-Y., CHUANG Y.-Y., CHEN B.-Y., OUHYOUNG M.: Video stabilization using robust feature trajectories. In *Proc. IEEE Intl. Conf. Computer Vision* (2009), pp. 1307–1404. 3

[LK81] LUCAS B. D., KANADE T.: An iterative image registration technique with an application to stereo vision. In *Proc. Intl. Joint Conf. Artificial Intelligence* (1981), pp. 674–679. 3

[LLM*11] LIN W.-Y., LIU S., MATSUSHITA Y., NG T.-T., CHEONG L. F.: Smoothly varying affine stitching. In *Proc. IEEE Intl. Conf. Computer Vision and Pattern Recognition* (2011), pp. 345–352. 3

[Low04] LOWE D. G.: Distinctive image features from scale-invariant keypoints. *Intl. Journal Computer Vision 60* (2004), 91–110. 6

[LSC*12] LUO S.-J., SHEN I.-C., CHEN B.-Y., CHENG W.-H., CHUANG Y.-Y.: Perspective-aware warping for seamless stereoscopic image cloning. *ACM Trans. Graphics (Proc. SIGGRAPH Asia 2012) 31*, 6 (2012), 182:1–182:8. 5

[OMGF*10] OLAVERRI-MONREAL C., GOMES P. E. R., FERNANDES R., VIEIRA F., FERREIRA M.: The see-through system: A vanet-enabled assistant for overtaking maneuvers. In *Proc. Intelligent Vehicles Sympos.* (2010), pp. 123–128. 3

[SS97] SZELISKI R., SHUM H.-Y.: Creating full view panoramic image mosaics and environment maps. In *Proc. ACM SIGGRAPH* (1997), pp. 251–258. 2, 3

[ST94] SHI J., TOMASI C.: Good feature to track. In *Proc. IEEE Intl. Conf. Computer Vision and Pattern Recognition* (1994), pp. 593–600. 6

[ST04] SAND P., TELLER S. J.: Video matching. *ACM Trans. Graphics 23*, 3 (2004), 592–599. 3

[TCGP09] TA D.-N., CHEN W.-C., GELDFAND N., PULLI K.: Surftrac: Efficient tracking and continuous object recognition using local feature descriptors. In *Proc. IEEE Intl. Conf. Computer Vision and Pattern Recognition* (2009), pp. 2937–2944. 3

[WFS*09] WANG Y.-S., FU H., SORKINE O., LEE T.-Y., SEIDEL H.-P.: Motion-aware temporal coherence for video resizing. *ACM Trans. Graphics 28*, 5 (2009). 3


**Figure 7:** *Qualitative comparisons. Challenging frames of reference sequence (a) and target sequence (b) are shown. (c) Aligned image using a global homography method. (d) Aligned image using Moving-DLT method [ZCBS13]. (e) Aligned image with our method.*

[WLY13] WU Y., LIM J., YANG M.-H.: Online object tracking: A benchmark. In *Proc. IEEE Intl. Conf. Computer Vision and Pattern Recognition* (2013), pp. 2411–2418. 4

[ZCBS13] ZARAGOZA J., CHIN T.-J., BROWN M. S., SUTER D.: As-projective-as-possible image stitching with moving dlt. In *Proc. IEEE Intl. Conf. Computer Vision and Pattern Recognition* (2013), pp. 2339–2346. 3, 6, 7, 9

[ZL14] ZHANG F., LIU F.: Parallax-tolerant image stitching. In *Proc. IEEE Intl. Conf. Computer Vision and Pattern Recognition* (2014). 6

[ZLY12] ZHONG W., LU H., YANG M.-H.: Robust object tracking via sparsity-based collaborative smodel. In *Proc. IEEE Intl. Conf. Computer Vision and Pattern Recognition* (2012), pp. 1838–1845. 4