June 2021

Object Detection and Tracking with Post-Merge Motion Estimation

Ruiduo Yang

Samuel William Hasinoff

Jinglun Gao

Donghui Han

Minghui Tan

Object Detection and Tracking with Post-Merge Motion Estimation

Abstract:

This publication describes techniques and apparatuses that enable an electronic device (e.g., a smartphone) with at least one camera to capture, render, and/or process frames. As the smartphone captures frames of a scene, the smartphone determines frames to be rendered and/or processed. Initially, the smartphone performs a global (fast) motion estimation to correctly render and/or process the captured frames of the scene. Information from the global motion estimation enables the smartphone to merge the captured frames, and the smartphone can generate one merged frame or a set of merged frame candidates. Information from the global motion estimation also enables the smartphone to perform target (object) detection on the merged frame. If the smartphone fails to detect a target (e.g., inside a target-detection box), the smartphone proceeds to render and/or process a next merged frame of the set of the merged frame candidates. If the smartphone, however, detects a target inside the target-detection box, the smartphone uses the necessary resources to perform local (accurate, detailed) motion estimation inside the target-detection box. The smartphone may use the information from the local motion estimation to merge the target, generate a local frame patch, and/or verify the target. The local motion estimation helps reduce or remove undesired artifacts (e.g., ghosting). Lastly, the smartphone projects the location of the target to a corresponding merged frame to generate a resulting frame with increased signal and/or increased signal-to-noise ratio (SNR).

Keywords:

Object detection, object tracking, photo, video, burst, frame, global motion estimation, local motion estimation, capturing, rendering, processing, merging, signal-to-noise ratio, SNR

Background:

Manufacturers of various electronic devices (e.g., smartphones, digital cameras) may embed one or more cameras in and/or on the electronic device that enable a user to capture static images (a single frame) or videos (multiple frames per second (FPS)). The quality of the captured frames, in part, depends on motion estimation that the smartphone performs under various lighting conditions, changing motion speeds of objects and/or targets in a scene, the motion of a user's hand holding the smartphone, and/or other factors, as is illustrated in Figure 1.

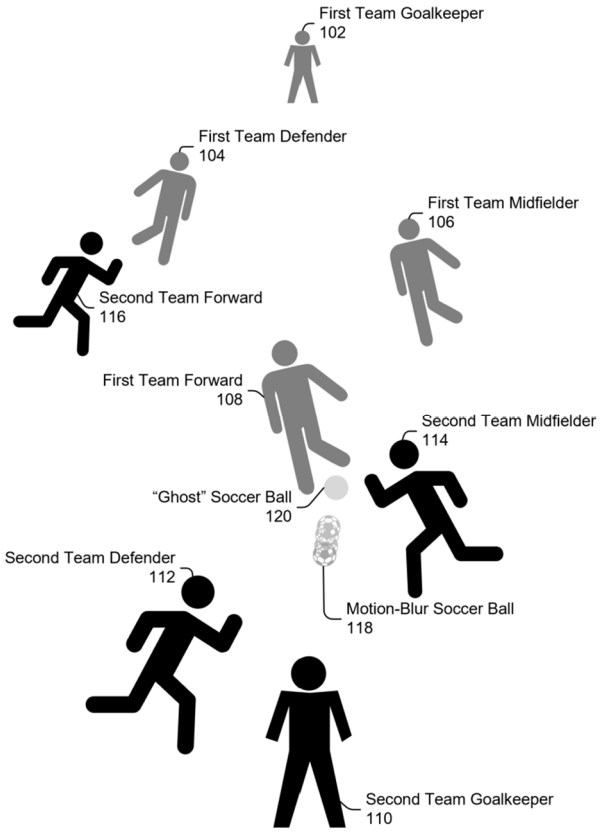

Figure 1

Figure 1 illustrates players of a soccer match between a first team (illustrated in grey) and a second team (illustrated in black). The first team includes at least a goalkeeper 102, a defender 104, a midfielder 106, and a forward 108. Similarly, the second team includes at least a goalkeeper 110, a defender 112, a midfielder 114, and a forward 116. Although not illustrated in Figure 1, the event may include additional players, fans who may be cheering and waving flags on the stands, referees in and/or around a pitch (soccer field), coaches, trainers, and/or journalists on the sidelines, and so forth. Also, in the pitch, during the illustrated soccer match, some players may be running, some players may be moving slowly or may be relatively stationary (e.g., the goalkeepers), and one of the players may be dribbling, passing, or striking the soccer ball. Also, the match may be played outdoors, and the lighting conditions may change due to movements of the clouds, time of the day, and/or other factors.

These factors may cause a smartphone to capture frames of the scene with undesired artifacts that may cause the smartphone to incorrectly determine a location of an object, for example, inside a target-detection box (not illustrated). Common undesired artifacts may be motion blur and/or "ghosting" of a target in the scene. For ease of description, in the same frame, Figure 1 illustrates an example motion-blur soccer ball 118 and an example "ghost" soccer ball 120 captured by the smartphone as the forward of the first team strikes the soccer ball. Motion blur may occur when a camera with a slow shutter speed captures a fast-moving target within a single frame. Ghosting may be an artifact when capturing a fast-moving target across multiple frames, and the camera fails to merge the multiple frames into a single frame correctly. In aspects, ghosting may appear as faded duplication(s) of the same fast-moving target in the scene.

One way to reduce these artifacts is to capture, render, and/or process the frames using shutter speeds greater than 30 frames per second (FPS) (e.g., 100 FPS). Capturing, rendering,

and/or processing frames using higher shutter speeds, however, is computationally expensive (e.g., image processing), consumes more power (e.g., battery power), requires additional volatile (e.g., dynamic random-access memory (DRAM)) and non-volatile (e.g., Flash) memory, and is difficult to share due to the required large files. Also, capturing, rendering, and/or processing frames using higher shutter speeds may result in a low signal and/or a low signal-to-noise ratio (SNR) of the captured frames.

Therefore, it is desirable to have a technological solution that can successfully capture, render, and/or process objects inside target-detection boxes with reduced undesired artifacts, increased signal, and/or increased SNR while conserving processing resources, power, and/or memory.

Description:

This publication describes techniques and apparatuses that enable an electronic device (e.g., a smartphone) with at least one camera to capture, render, and/or process frames with reduced undesired artifacts (e.g., ghosting), increased signal, and/or increased signal-to-noise ratio (SNR). The techniques and apparatus described herein can reduce the undesired artifacts, increase the signal, and/or increase the SNR without requiring additional processing resources, consuming more power, using more memory, and/or using a more advanced optical technology (e.g., complementary metal-oxide-semiconductor (CMOS) image sensor, charge-coupled device (CCD) image sensor).

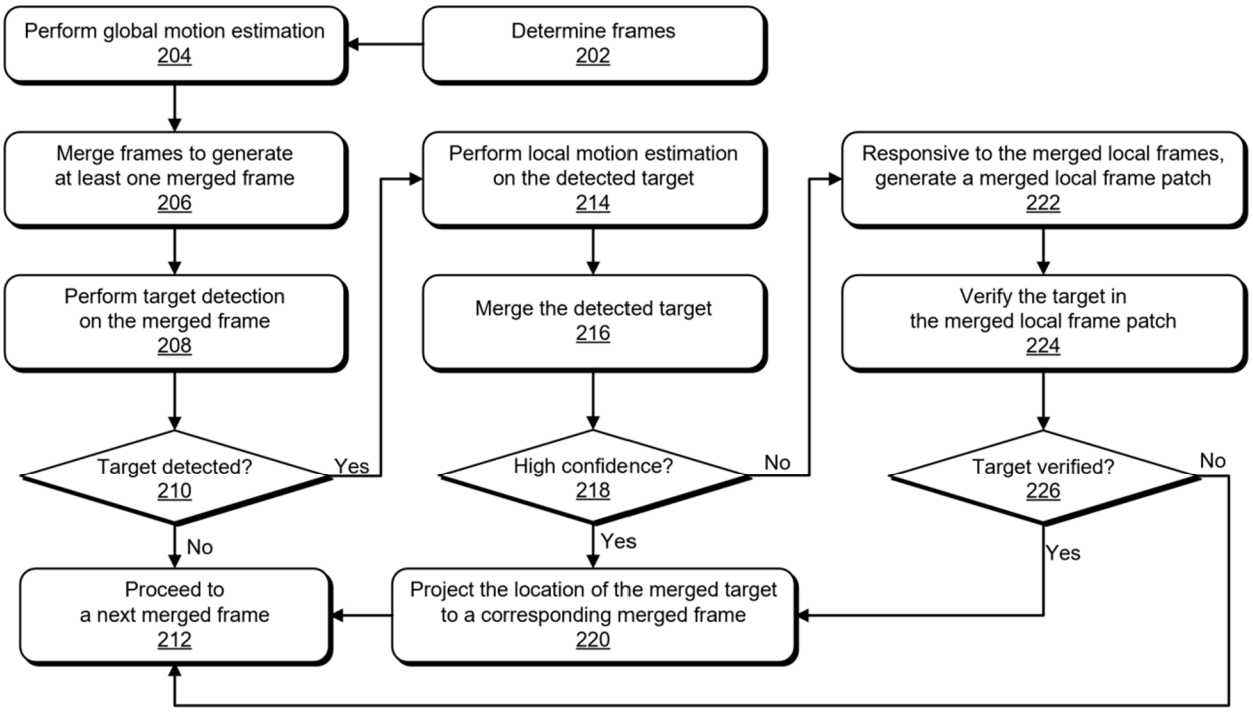

Figure 2

Figure 2 depicts a block diagram illustrating a technique implemented on a smartphone for capturing, rendering, and/or processing frames to produce images with reduced undesired artifacts, increased signal, and/or increased SNR. As the smartphone captures (takes a photograph, takes a burst of photographs, records a video) frames of a scene, at stage 202, the smartphone determines frames to be rendered and/or processed. The scene may include moving and/or stationary targets (objects).

At stage 204, the smartphone performs a global (fast) motion estimation of the targets in the scene and renders and/or processes the captured frames by using various techniques and/or components embedded in and/or on the smartphone. For example, the smartphone may utilize an auto-exposure (AE) algorithm. The auto-exposure algorithm may combine gyroscope data, accelerometer data, image sensor (e.g., CMOS image sensor) data, ambient-light sensor data,

luminance statistics, and so forth to determine the movement (or lack thereof) of the target(s) in the scene.

In the context of the scene illustrated in Figure 1, at stage 204 of Figure 2, the smartphone performs the global motion estimation to determine the motion of the players, the soccer ball, fans, flags, referees, coaches, journalists, and so forth. In one aspect, the global motion estimation may accurately determine the movement of the players, fans, flags, referees, coaches, journalists, and so forth, but may fail to accurately determine the movement of the soccer ball due to overcasting and low-light conditions, as is illustrated in Figure 1. As another example (not illustrated), an autonomous-driving and/or driver-assistance vehicle may utilize the techniques described herein to detect and avoid colliding with objects. The vehicle may be driving through a tunnel with low-lighting conditions. As the vehicle drives near an exit of the tunnel, light from the Sun may cause severe backlighting of objects adjacent and/or in front of the vehicle. The vehicle may accurately determine movements of other vehicles but, without further processing, the vehicle may fail to accurately determine a movement of a wild animal (e.g., a raccoon, a deer) who may have wandered in the tunnel. As such, the global motion estimation may be considered a first-order motion estimation. The techniques described herein, however, perform further processing to accurately determine locations and/or movements of objects in various lighting conditions, as is further described below.

At stage 206, information from the global motion estimation enables the smartphone to merge the captured frames that better represent the scene. The smartphone can generate one merged frame or a set of merged frame candidates. Generating a set of merged frame candidates may be useful to narrow down a target of interest. To merge the frames, the smartphone may use various techniques based, in part, on performance requirements and/or applications. In one aspect,

to merge the frames, the smartphone uses information from the global motion estimation to align the frames that represent the scene better (e.g., the whole scene in Figure 1) and generate the merged frame with a high signal and/or a high SNR.

The high signal and/or the high SNR helps the smartphone to perform target (object) detection and/or target tracking with considerable accuracy. As such, after the smartphone generates the merged frame (or a set of merged frame candidates), at stage 208, the smartphone performs over-detection for the target detection and/or the target tracking on the merged frame. In one aspect, the smartphone may use a box around the detected target (target-detection box). In general, the target-detection box is smaller than the merged frame. Also, the target-detection box has a lower resolution than the merged frame. Therefore, the smartphone can perform the over-detection because rendering and/or processing target-detection boxes requires fewer computational resources, less power, and/or less memory, and/or may be achieved in a shorter amount of time compared to the merged frame (e.g., the whole scene in Figure 1).

At stage 210, the smartphone determines whether there is a target inside a target-detection box. If the smartphone fails to detect a target, at stage 212, the smartphone proceeds to a next merged frame of the set of merged frame candidates. If the smartphone detects a target inside a target-detection box, at stage 214, the smartphone performs a local (accurate, detailed) motion estimation on the detected target. Continuing with the example of Figure 1, if the smartphone detects the soccer ball inside a target-detection box, the smartphone performs a local motion estimation of the soccer ball. At stage 214, the smartphone can use the necessary computational resources, power, and/or memory to perform local motion estimation of the soccer ball inside the target-detection box instead of the whole merged frame (e.g., the whole scene in Figure 1).

At stage 216, responsive to the local motion estimation with accurate and detailed motion estimation of the target, the smartphone merges the detected target inside the target-detection box. The smartphone then, at stage 218, computes a confidence score on the merged detected target. If the smartphone determines with high confidence that the merged detected target is accurate and sharp (e.g., no ghosting), the smartphone then, at stage 220, projects the location of the merged target (e.g., the soccer ball) to a corresponding merged frame and reports the location of the target in a resulting frame. Consequently, the smartphone can accurately determine the position of the soccer ball and reduces or eliminates undesired artifacts (e.g., without a "ghost" soccer ball 120) in the resulting frame. After reporting the location of the target in the resulting frame, the smartphone proceeds to a next merged frame (stage 212) of the set of the merged frame candidates.

If the smartphone, however, at stage 218 fails to determine with high confidence the merged target, at stage 222, the smartphone generates a merged local frame patch based on merged local frames. The merged local frame patch may include a higher signal, a higher SNR, and/or a sharper image of the target. The sharper image of the target, at stage 224, enables the smartphone to verify the target in the merged local frame patch. Verifying the target includes distinguishing the target from an undesired artifact (e.g., ghosting) produced by capturing, rendering, and/or processing the target.

Verifying the target also includes verifying the target as being a target of interest. The target of interest, however, may vary depending on an application that utilizes target detection. Continuing with the example of the autonomous-driving and/or driver-assistance vehicle, verifying the target of interest may be important for safety reasons. The vehicle determines whether the target of interest is a pedestrian, a pet, another vehicle, a barrier, and so forth to enable safe driving. The vehicle may perform a different driving maneuver when verifying the target as

a pedestrian compared to verifying the target as a tumbleweed, trash (e.g., a plastic bag), or a soccer ball. The vehicle avoids colliding with the pedestrian, but the vehicle, depending on the situation (e.g., high speeds), may drive through the plastic bag. Continuing with the example of Figure 1, a soccer fan using a smartphone to capture, render, and/or process the scene in Figure 1 requires the smartphone to verify that the soccer ball is a target of interest. Therefore, depending on the application, an electronic device utilizing the techniques described herein may use various parameters, methods, sensors, and so forth to verify whether a target is a target of interest.

At stage 226, the smartphone determines whether a target is verified in the merged local frame patch. If the smartphone fails to verify a target in the merged local frame patch, the smartphone proceeds to a next merged frame (stage 212) of the set of the merged frame candidates. If, however, the smartphone determines that the merged local frame patch includes a verified target, the smartphone projects the location of the target to a corresponding merged frame and generates a resulting frame. The techniques described in Figure 2 are an iterative process and can be used in real-time applications. Thus, next, at stage 212, the smartphone proceeds to a next merged frame of the set of merged frame candidates generated at stage 206.

References:

[1] Patent Publication: US20170053167A1. Ren et al. Systems and Methods for Object Tracking. Priority Date: August 18, 2015.

[2] Patent Publication: US20130177203A1. Koo et al. Object Tracking and Processing. Priority Date: January 6, 2012.

[3] Patent Publication: US20120154579A1. Hampapur et al. Detection and Tracking of Moving Objects. Priority Date: December 20, 2010.

[4] Patent Publication: US20160248979A1. Israel et al. Digital Image Processing. Priority Date: July 23, 2013.