Deep Spatial Gradient and Temporal Depth Learning for Face Anti-spoofing (Appendix)

1. Temporal Depth in Face Anti-spoofing

In this section, we use some simple examples to explain that exploiting temporal depth and motion is reasonable in the face anti-spoofing task.

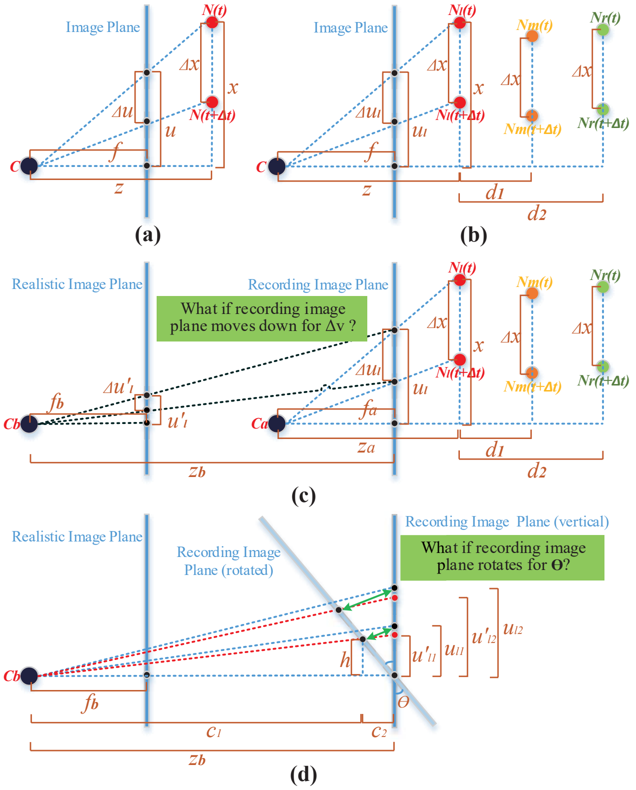

Figure 1. The schematic diagram of motion and depth variation in different scenes.

1.1. Basic Scene

As shown in Fig. 1(a), node C denotes the camera focus. **Image Plane** represents the image plane of camera. $N(t)$ is one facial point at time t, and $N(t+\Delta t)$ is the corresponding point when $N(t)$ moves down vertically for Δx at time $t+\Delta t$. For example, $N(t)$ can be the point of nose or ear. f denotes the focal distance, and z is the horizontal distance from the focal point to the point $N(t)$. u and x are the corresponding coordinates in vertical dimension. When $N(t)$ moves down vertically to $N(t+\Delta t)$ for Δx, the motion can be reflected on the image plane as Δu. According to the camera model, we can obtain:

$$\begin{aligned} \frac{x}{u} &= \frac{z}{f}, \\ \Leftrightarrow u &= \frac{fx}{z}. \end{aligned} \tag{1}$$

When $N(t)$ moves down vertically for Δx to $N(t+\Delta t)$, the Δu can be achieved:

$$\Delta u = \frac{f \Delta x}{z}. \tag{2}$$

As shown in Fig. 1(b), to distinguish points N_l, N_m and N_r, we transform Eq. 2 and get Δu_l, Δu_m and Δu_r (Δu_m and Δu_r are not shown in the figure):

$$\begin{aligned} \Delta u_l &= \frac{f \Delta x}{z}, \\ \Delta u_m &= \frac{f \Delta x}{z + d_1}, \\ \Delta u_r &= \frac{f \Delta x}{z + d_2}, \end{aligned} \tag{3}$$

where d_1 and d_2 are the corresponding depth difference. From Eq. 3, there are:

$$\begin{aligned} \frac{\Delta u_l}{\Delta u_m} &= \frac{z + d_1}{z} = \frac{d_1}{z} + 1, \\ \frac{\Delta u_l}{\Delta u_r} &= \frac{z + d_2}{z} = \frac{d_2}{z} + 1. \end{aligned} \tag{4}$$

Removing z from Eq. 4, d_1/d_2 can be obtained:

$$\frac{d_1}{d_2} = \frac{\frac{\Delta u_l}{\Delta u_m} - 1}{\frac{\Delta u_l}{\Delta u_r} - 1}, \tag{5}$$

In this equation, we can see that the relative depth d_1/d_2 can be estimated by the motion of three points, when $d_2 \neq 0$. The equations above are about the real scenes. In the following, we will introduce the derivation of attack scenes.

1.2. Attack Scene

1.2.1 What if the attack carriers move?

As shown in Fig. 1(c), there are two image spaces in attack scenes: one is recording image space, where we replace z, f by z_a, f_a, and the other is realistic image space, where we replace z, f by z_b, f_b. In the recording image space, it's similar to Eq. 3:

$$\begin{aligned} \Delta u_l &= \frac{f_a \Delta x}{z_a}, \\ \Delta u_m &= \frac{f_a \Delta x}{z_a + d_1}, \\ \Delta u_r &= \frac{f_a \Delta x}{z_a + d_2}, \end{aligned} \tag{6}$$

where $\Delta u_l, \Delta u_m, \Delta u_r$ are the magnitude of optical flow when three points $N_l(t), N_m(t), N_r(t)$ move down vertically for Δx.

In the realistic image space, there are:

$$\begin{aligned} \Delta u_l' &= \frac{f_b \Delta x_l}{z_b}, \\ \Delta u_m' &= \frac{f_b \Delta x_m}{z_b}, \\ \Delta u_r' &= \frac{f_b \Delta x_r}{z_b}, \end{aligned} \tag{7}$$

where Δx_l, Δx_m and Δx_r are the motion of three points on the recording image plane, and $\Delta u_l, \Delta u_m, \Delta u_r$ are the corresponding values mapping on the realistic image plane.

Actually, there are $\Delta x_l = \Delta u_l, \Delta x_m = \Delta u_m, \Delta x_r = \Delta u_r$, if the recording screen is static. Now, a vertical motion Δv is given to the recording screen, just as $\Delta x_l = \Delta u_l + \Delta v, \Delta x_m = \Delta u_m + \Delta v, \Delta x_r = \Delta u_r + \Delta v$. By inserting Δv, we transform Eq. 7 into:

$$\begin{aligned} \Delta u_l' &= \frac{f_a f_b \Delta x + z_a f_b \Delta v}{z_a z_b}, \\ \Delta u_m' &= \frac{f_a f_b \Delta x + (z_a + d_1) f_b \Delta v}{(z_a + d_1) z_b}, \\ \Delta u_r' &= \frac{f_a f_b \Delta x + (z_a + d_2) f_b \Delta v}{(z_a + d_2) z_b}, \end{aligned} \tag{8}$$

Due to that only $\Delta u_l', \Delta u_m', \Delta u_r'$ can be observed directly in the sequential images, we can estimate the relative depth via $\Delta u_l', \Delta u_m', \Delta u_r'$. So we leverage Eq. 5 to estimate the relative depth d_1'/d_2':

$$\frac{d_1'}{d_2'} = \frac{\dfrac{\Delta u_l'}{\Delta u_m'} - 1}{\dfrac{\Delta u_l'}{\Delta u_r'} - 1}, \tag{9}$$

and then we can insert Eq. 8 into Eq. 9 to get:

$$\frac{d_1'}{d_2'} = \frac{d_1}{d_2} \cdot \frac{f_a \Delta x + (z_a + d_2) \Delta v}{f_a \Delta x + (z_a + d_1) \Delta v}. \tag{10}$$

According to equations above, some important conclusions can be summarized:

- If $\Delta x = 0$, the scene can be recognized as print attack and Eq. 10 will be invalid, for $\Delta u_l' = \Delta u_r'$, and the denominator in Eq. 9 will be zero. So here we use Eq. 8 and

  $$\begin{aligned} \frac{\Delta u_l'}{\Delta u_m'} &= \frac{d_1'}{z_b} + 1, \\ \frac{\Delta u_l'}{\Delta u_r'} &= \frac{d_2'}{z_b} + 1, \end{aligned} \tag{11}$$

  to obtain:

  $$d_1' = d_2' = 0. \tag{12}$$

  In this case, it's obvious that the facial relative depth is abnormal and the face is fake.

- If $\Delta x \neq 0$, the scene can be recognized as replay attack.

  - If $\Delta v = 0$, there is:

    $$\frac{d_1'}{d_2'} = \frac{d_1}{d_2}. \tag{13}$$

    In this case, if these two image planes are parallel and the single-frame model can not detect the static spoof cues, the model will fail in the task of face anti-spoofing, owing to that the model is hard to find the abnormality of relative depth estimated from the facial motion. We call this scene **Perfect Spoofing Scene(PSS)**. Of course, making up **PSS** will cost a lot and is approximately impossible in practice.

  - If $\Delta v \neq 0$ and we want to meet Eq. 13, the following equation should be satisfied:

    $$\frac{f_a \Delta x + (z_a + d_2) \Delta v}{f_a \Delta x + (z_a + d_1) \Delta v} = 1, \tag{14}$$

    then,

    $$\begin{aligned} &(d_2 - d_1) \Delta v = 0, \\ \Leftrightarrow &d_2 - d_1 = 0, \; if \; \Delta v \neq 0. \end{aligned} \tag{15}$$

    However, in our assumption, $d_1 \neq d_2$, so:

    $$\frac{d_1'}{d_2'} \neq \frac{d_1}{d_2}. \tag{16}$$

This equation indicates that relative depth can't be estimated preciously, if the attack carrier moves in the replay attack. And Δv usually varies when attack carrier moves in the long-term sequence, leading to the variation of d'_1/d'_2. This kind of abnormality is more obvious along with the long-term motion.

- If d_2 denotes the largest depth difference among facial points, then $d_1/d_2 \in [0, 1]$, showing that constraining depth label of living face to $[0, 1]$ is valid. As analyzed above, for spoofing scenes, the abnormal relative depth usually varies over time, so it is too complex to be computed directly. Therefore, we merely set depth label of spoofing face to all 0 to distinguish it from living label, making the model learn the abnormity under depth supervision itself.

1.2.2 What if the attack carriers rotate?

As shown in Fig. 1(d), we rotate the recording image plane for degree θ. u_{l2}, u_{l1} are the coordinates of $N_l(t), N_l(t + \Delta t)$ mapping on the recording image plane. The two black points at the *right* end of green double arrows on recording image plane (vertical) will reach the two black points at the *left* end of green double arrow on recording image plane (rotated), when the recording image plane rotates. And the corresponding values u_{l2}, u_{l1} will not change after rotation. For convenient computation, we still map the rotated points to the vertical recording image plane. And the coordinates after mapping are u'_{l2}, u'_{l1}. c_1, c_2, h are the corresponding distances shown in the figure. According to the relationship of the foundamental variables, we can obtain:

$$\begin{aligned} h &= u_{l1} \cos\theta, \\ \frac{z_b}{c_1} &= \frac{u'_{l1}}{h}, \\ c_2 &= u_{l1} \sin\theta, \\ c_1 + c_2 &= z_b. \end{aligned} \tag{17}$$

Deriving from equations above, we can get u'_{l1}:

$$u'_{l1} = \frac{z_b u_{l1} \cos\theta}{z_b - u_{l1} \sin\theta}, \tag{18}$$

and u'_{l2} can also be calculated by imitating Eq. 18:

$$u'_{l2} = \frac{z_b u_{l2} \cos\theta}{z_b - u_{l2} \sin\theta}. \tag{19}$$

Subtract u'_{l1} from u'_{l2}, the following is achieved:

$$u'_{l2} - u'_{l1} = (u_{l2} - u_{l1}) \cdot \frac{{z_b}^2 \cos\theta}{(z_b - u_{l1} \sin\theta)(z_b - u_{l2} \sin\theta)}. \tag{20}$$

Obviously, $u_{l2} - u_{l1} = \Delta u_l$. We define $u'_{l2} - u'_{l1} = \Delta u_l^\theta$. And then we get the following equation:

$$\begin{aligned} \Delta u_l^\theta &= \Delta u_l \cdot \frac{{z_b}^2 \cos\theta}{(z_b - u_{l1} \sin\theta)[z_b - (u_{l1} + \Delta u_l) \sin\theta]}, \\ \Delta u_m^\theta &= \Delta u_m \cdot \frac{{z_b}^2 \cos\theta}{(z_b - u_{m1} \sin\theta)[z_b - (u_{m1} + \Delta u_l) \sin\theta]}, \\ \Delta u_r^\theta &= \Delta u_r \cdot \frac{{z_b}^2 \cos\theta}{(z_b - u_{r1} \sin\theta)[z_b - (u_{r1} + \Delta u_l) \sin\theta]}, \end{aligned} \tag{21}$$

where the relationship between $\Delta u_m^\theta, \Delta u_r^\theta$ and $N_m(t), N_r(t)$ are just like that between Δu_l^θ and $N_l(t)$, as well as u_{m1}, u_{r1}. Note that for simplification, we only discuss the situation that u_{l1}, u_{m1}, u_{r1} are all positive.

Reviewing Eq. 7, We can confirm that $\Delta x_l = \Delta u_l^\theta$, $\Delta x_m = \Delta u_m^\theta$, $\Delta x_r = \Delta u_r^\theta$. According to Eq. 9, the final d'_1/d'_2 can be estimated:

$$\frac{d'_1}{d'_2} = \frac{\dfrac{\Delta u_l}{\Delta u_m} \cdot \beta_1 - 1}{\dfrac{\Delta u_l}{\Delta u_r} \cdot \beta_2 - 1} = \frac{(\dfrac{d_1}{z_a} + 1) \cdot \beta_1 - 1}{(\dfrac{d_2}{z_a} + 1) \cdot \beta_2 - 1}, \tag{22}$$

where β_1 and β_2 can be represented as:

$$\begin{aligned} \beta_1 &= \frac{(z_b - u_{m1} \sin\theta)(z_b - u_{m2} \sin\theta)}{(z_b - u_{l1} \sin\theta)(z_b - u_{l2} \sin\theta)}, \\ \beta_2 &= \frac{(z_b - u_{r1} \sin\theta)(z_b - u_{r2} \sin\theta)}{(z_b - u_{l1} \sin\theta)(z_b - u_{l2} \sin\theta)}, \end{aligned} \tag{23}$$

where $u_{l2} = u_{l1} + \Delta u_l, u_{m2} = u_{m1} + \Delta u_m, u_{r2} = u_{r1} + \Delta u_r$. Observing Eq. 22, we can see if $\beta_1 < 1, \beta_2 > 1$ or $\beta_1 > 1, \beta_2 < 1$, there will be $d'_1/d'_2 \neq d_1/d_2$, .

Now, we discuss the sufficient condition of $\beta_1 < 1, \beta_2 > 1$. When $u_{m1} > u_{l1}, u_{m2} > u_{l2}, u_{r1} < u_{l1}, u_{r2} < u_{l2}$, the $\beta_1 < 1, \beta_2 > 1$ can be established. Similar to Eq. 3, the relationship of variables can be achieved:

$$\begin{aligned} &\frac{f_a x_{l1}}{z_a} = u_{l1}, \frac{f_a x_{l2}}{z_a} = u_{l2}, \\ &\frac{f_a x_{m1}}{z_a + d_1} = u_{m1}, \frac{f_a x_{m2}}{z_a + d_1} = u_{m2}, \\ &\frac{f_a x_{r1}}{z_a + d_2} = u_{r1}, \frac{f_a x_{r2}}{z_a + d_2} = u_{r2}, \end{aligned} \tag{24}$$

From Eq. 24 and $u_{m1} > u_{l1}, u_{m2} > u_{l2}$, we can obtain:

$$\begin{aligned} x_{m1} &> x_{l1} \cdot \frac{z_a + d_1}{z_a}, \\ x_{m1} + \Delta x &> (x_{l1} + \Delta x) \cdot \frac{z_a + d_1}{z_a}, \end{aligned} \tag{25}$$

$$\begin{aligned} x_{r1} &< x_{l1} \cdot \frac{z_a + d_2}{z_a}, \\ x_{r1} + \Delta x &< (x_{l1} + \Delta x) \cdot \frac{z_a + d_2}{z_a}, \end{aligned} \tag{26}$$

where $x_{l1}, x_{l2}, x_{m1}, x_{m2}, x_{r1}, x_{r2}$ are corresponding coordinates of $N_l(t+\Delta t), N_l(t), N_m(t+\Delta t), N_m(t), N_r(t+\Delta t), N_r(t)$ in the dimension of **x** in recording image space. In facial regions, we can easily find corresponding points $N_l(t), N_m(t)$, which satisfy that $d_1 \ll z_a$ (i.e., $d_1 = 0$) and $x_{m1} > x_{l1}$. In this pattern, Eq. 25 can be established. To establish Eq. 26, we only need to find point $N_r(t)$, which satisfies that $x_{r1} < x_{l1}$. According to the derivation above, we can see that there exists cases that $d_1'/d_2' < d_1/d_2$. And there are also many cases that satisfy $d_1'/d_2' > d_1/d_2$, which we do not elaborate here. When faces move, the absolute coordinates $x_{l1}, x_{l2}, x_{m1}, x_{m2}, x_{r1}, x_{r2}$ vary, as well as β_1, β_2, leading to the variation of estimated relative depth of three facial points at different moments, which will not occur in the *real* scene. That's to say, if the realistic image plane and recording image plane are not parallel, we can seek cases to detect abnormal relative depth with the help of abnormal facial motion.

1.2.3 Discussion

One of basis of the elaboration above is that the structure of face is similar to that of the hill, which is complex, dense and undulate. This is interesting and worth being exploited in face anti-spoofing.

Even though we only use some special examples to demonstrate our viewpoints and assumption, they can still prove the reasonability of utilizing facial motion to estimate the relative facial depth in face anti-spoofing task. In this way, the learned model can seek the abnormal relative depth and motion in the facial regions. And our extensive experiment demonstrates our assumption and indicates that temporal depth method indeed improves the performance of face anti-spoofing.