

Now you see me, now you don't

Ataki na systemy przetwarzania obrazu na przykładzie HOG

Magdalena Mozgawa

WMI UAM

1 czerwca 2020

1 Systemy przetwarzania obrazu w życiu codziennym

- Czym jest obraz?
- Przetwarzanie obrazu. Definicja i zastosowania

2 Sposoby wykrywania obiektów (twarzy) na zdjęciach

- Popularne metody
- Histogramy zorientowanych gradientów: omówienie

3 Atak: jak zdezorientować gradienty?

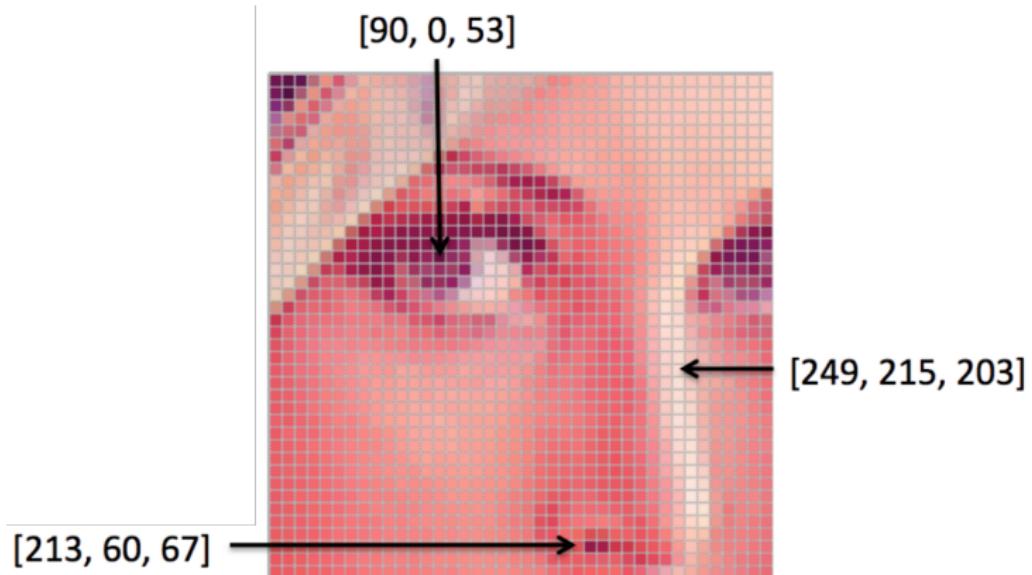
4 Podsumowanie

- Wnioski
- Źródła

Obraz jako dwuwymiarowy rzut trójwymiarowej rzeczywistości



Jak zapisać obraz?



Jak zapisać obraz?

$$\begin{matrix} & 1 & 2 & \dots & n \\ \begin{matrix} 1 \\ 2 \\ 3 \\ \vdots \\ m \end{matrix} & \left[\begin{matrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ a_{31} & a_{32} & \dots & a_{3n} \\ \vdots & \vdots & \vdots & \vdots \\ a_{m1} & a_{m2} & \dots & a_{mn} \end{matrix} \right] \end{matrix}$$

Czym jest przetwarzanie obrazu

Przetwarzanie obrazu (computer vision): dziedzina zajmująca się rozwojem technik pozwalających na odzyskiwanie ze zdjęć trójwymiarowego wyglądu obiektów [A1]

Przetwarzanie obrazu w służbie człowiekowi

- diagnostyka medyczna
- Optical Character Recognition i Optical Mark Recognition
- autonomiczne pojazdy
- wspomaganie osób z niepełnosprawnościami
- rozpoznawanie dźwięku
- w dronach - patrolowanie zbiorów, poszukiwania ludzi w pożarach, zawalonych budynkach, itp [A1]

Co można zrobić z Twoim zdjęciem



S

Co można zrobić z Twoim zdjęciem



Co można zrobić z Twoim zdjęciem



Clearview AI has expanded to at least 26 countries outside the US, engaging national law enforcement agencies, government bodies, and police forces in Australia, Belgium, Brazil, Canada, Denmark, Finland, France, Ireland, India, Italy, Latvia, Lithuania, Malta, the Netherlands, Norway, Portugal, Serbia, Slovenia, Spain, Sweden, Switzerland, and the United Kingdom. [B2]

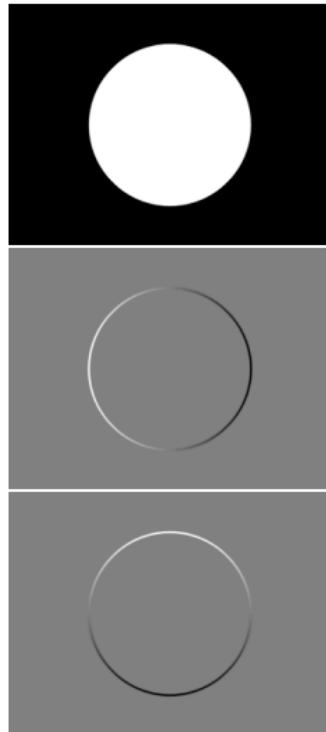
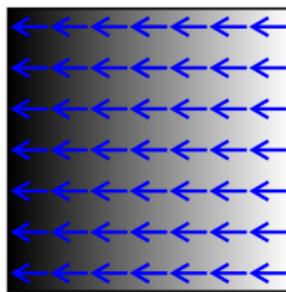
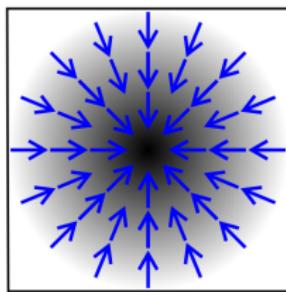
- 1 Systemy przetwarzania obrazu w życiu codziennym
 - Czym jest obraz?
 - Przetwarzanie obrazu. Definicja i zastosowania
- 2 Sposoby wykrywania obiektów (twarzy) na zdjęciach
 - Popularne metody
 - Histogramy zorientowanych gradientów: omówienie
- 3 Atak: jak zdezorientować gradienty?
- 4 Podsumowanie
 - Wnioski
 - Źródła

Popularne metody

- Algorytm Viola-Jonesa
- Histogram zorientowanych gradientów
- Głębokie uczenie maszynowe, w szczególności konwolucyjne sieci neuronowe (CNNs)



Gradient i zorientowane gradienty

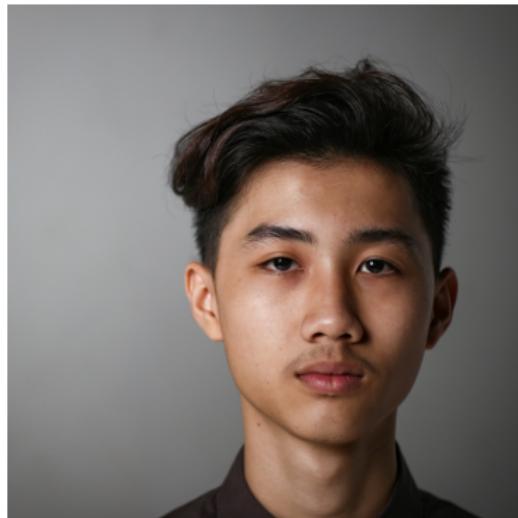


Wielkość wektora gradientu:

$$\sqrt{(f_x^2 + f_y^2)}.$$

Kąt wektora gradientu: $\arctan \frac{f_y}{f_x}$.

Zorientowane gradienty: przykład



$$\begin{pmatrix} 128 & 136 & 119 \\ 142 & 132 & 147 \\ 136 & 148 & 161 \end{pmatrix}$$

$$f_x = 147 - 142$$

$$f_y = 148 - 136$$

Wielkość wektora gradientu:
 $\sqrt{(5^2 + 12^2)} = 13.$

Kąt wektora gradientu:
 $\arctan 5/12 \approx 23^\circ.$

Zorientowane gradienty: przykład



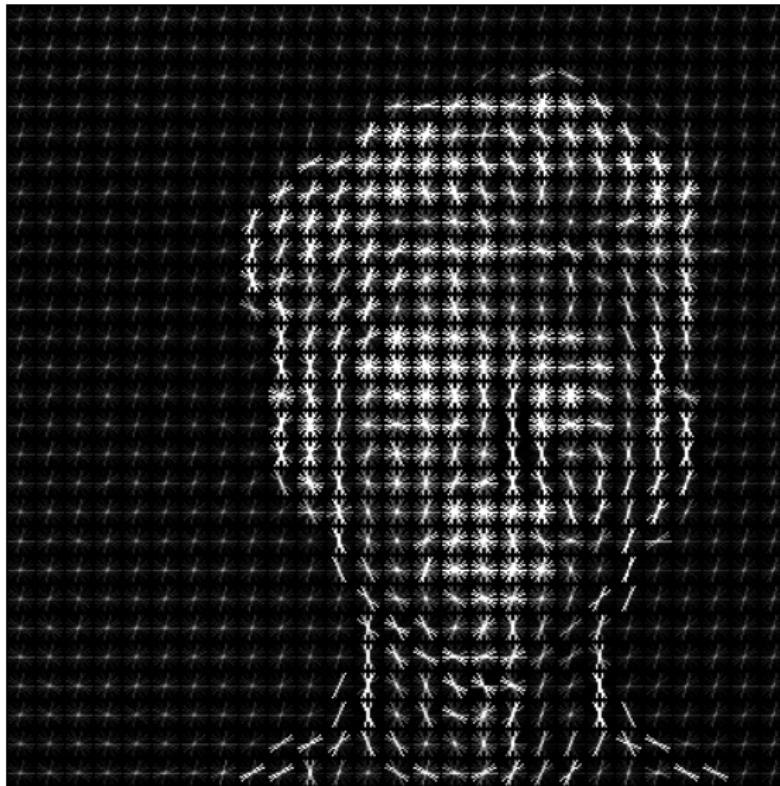
Histogram zorientowanych gradientów

Wielkości wektorów: $\begin{pmatrix} 25 & 18 \\ 15 & 10 \end{pmatrix}$, kąty wektorów: $\begin{pmatrix} 47 & 172 \\ 42 & 50 \end{pmatrix}$

0	20	40	60	80	100	120	140	160
0	0	0	0	0	0	0	0	0
0	0	25	0	0	0	0	0	0
18	0	25	0	0	0	0	0	0
18	0	40	0	0	0	0	0	0
18	0	45	5	0	0	0	0	0

Tym sposobem z $8 \times 8 \times 3 \Rightarrow 8 \times 8 \times 2 \Rightarrow 9$ wartości opisujących podobraz.

Histogram zorientowanych gradientów: przykład

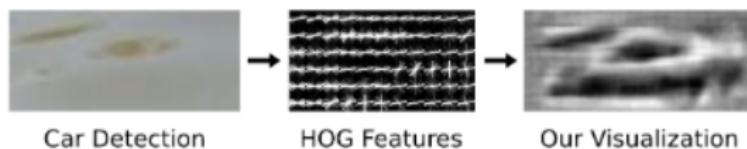


- 1 Systemy przetwarzania obrazu w życiu codziennym
 - Czym jest obraz?
 - Przetwarzanie obrazu. Definicja i zastosowania
- 2 Sposoby wykrywania obiektów (twarzy) na zdjęciach
 - Popularne metody
 - Histogramy zorientowanych gradientów: omówienie
- 3 Atak: jak zdezorientować gradienty?
- 4 Podsumowanie
 - Wnioski
 - Źródła

Czasami gradienty same się dezorientują



Figure 1: An image from PASCAL and a high scoring car detection from DPM [8]. Why did the detector fail?



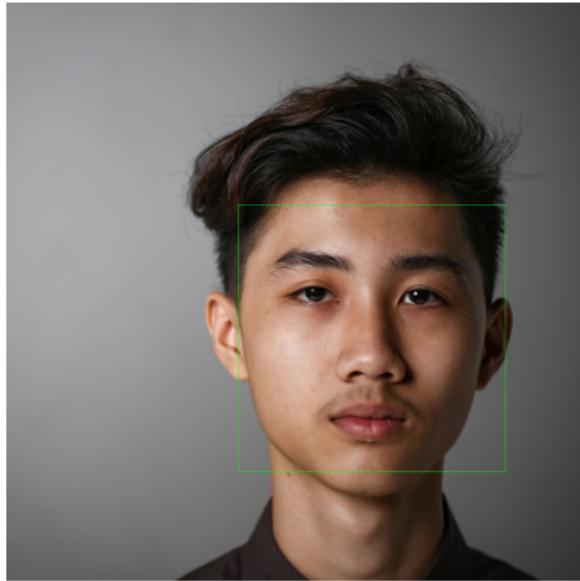
Rysunek: If it looks like a car and it honks like a car... [A3]

Twarz klasyfikatora

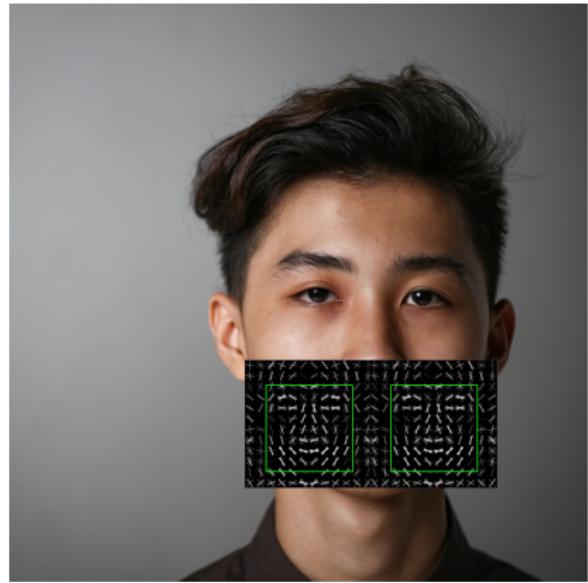
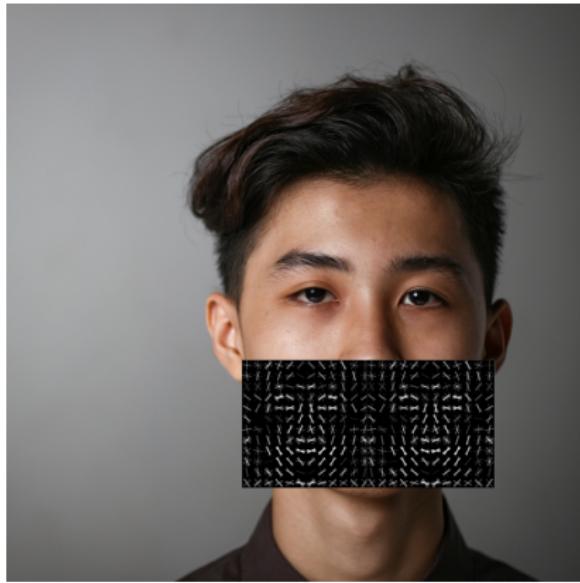


Rysunek: HOG: wizualizacja 'twarzy człowieka' z biblioteki dlib [C2]

Rozpoznawanie twarzy z biblioteką dlib



Rozpoznawanie twarzy z biblioteką dlib



Zastosowanie w praktyce



- 1 Systemy przetwarzania obrazu w życiu codziennym
 - Czym jest obraz?
 - Przetwarzanie obrazu. Definicja i zastosowania
- 2 Sposoby wykrywania obiektów (twarzy) na zdjęciach
 - Popularne metody
 - Histogramy zorientowanych gradientów: omówienie
- 3 Atak: jak zdezorientować gradienty?
- 4 Podsumowanie
 - Wnioski
 - Źródła

Wnioski

- HOG jest dość starą (2005) metodą wykrywania obiektów na zdjęciach, ale jest jednocześnie bardzo łatwy w implementacji (OpenCV + dlib)
- Bez względu na użytą metodę wykrywania, systemy są podatne na ataki i można to wykorzystać w różnych celach:
 - ▶ oszukiwanie systemów wykrywania twarzy [A6, A7]
 - ▶ modyfikowanie znaków drogowych w celu oszukania maszyn [A5, B3]
 - ▶ podrabianie podpisów tradycyjnych weryfikowanych elektronicznie [A4]
- Wszystkie Wasze zdjęcia, które postujecie w mediach społecznościowych są publiczne i zostaną użyte bez Waszej wiedzy i zgody do wytrenowania jakichś modeli

Źródła naukowe (A)

- ① Szeliski, R. Computer Vision: Algorithms and Applications. Springer: 2010.
- ② Dalal, N., Triggs, B. Histogram of Oriented Gradients for Human Detection. [In proceedings] IEEE Computer Society Conference on Computer Vision and Pattern Recognition: 2005.
- ③ Vondrick C. et al. HOGgles: Visualizing Object Detection Features. [In proceedings] IEEE International Conference on Computer Vision: 2013.
- ④ Hafemann, L., Sabourin, R. Characterizing and evaluating adversarial examples for Offline Handwritten Signature Verification.
<https://arxiv.org/pdf/1901.03398.pdf>
- ⑤ Eykholt, K. et al. Robust Physical-World Attacks on Deep Learning Visual Classification. [In proceedings] IEEE/CVF Conference on Computer Vision and Pattern Recognition: 2018.
- ⑥ Sharif, M. et al. Accessorize to a Crime: Real and Stealthy Attack on State-of-the-Art Face Recognition. CCS '16: Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security: 2016.
- ⑦ Komkov, S., Petiushko, A. AdvHat: Real-world adversarial attack on AdrFace Face ID system. 2019. <https://arxiv.org/abs/1908.08705>

Źródła prasowe (B)

- ① The Secretive Company That Might End Privacy as We Know It
<https://www.nytimes.com/2020/01/18/technology/clearview-privacy-facial-recognition.html>
- ② Clearview's Facial Recognition App Has Been Used By The Justice Department, ICE, Macy's, Walmart, And The NBA
<https://www.buzzfeednews.com/article/ryanmac/clearview-ai-fbi-ice-global-law-enforcement>
- ③ Model Hacking ADAs to Pave Safer Roads for Autonomous Vehicles
<https://www.mcafee.com/blogs/other-blogs/mcafee-labs/model-hacking-adas-to-pave-safer-roads-for-autonomous-veh>

Repozytoria kodu (C)

- ① Adversarial-Faces by BruceMacD
<https://github.com/BruceMacD/Adversarial-Faces>
- ② Dlib 18.6 released: Make your own object detector!
<http://blog.dlib.net/2014/02/dlib-186-released-make-your-own-object.html>

Źródła obrazów

- ① Wikipedia (Hasła: Lenna, Macierz, ClearView AI)
- ② Stanford Introduction to Computer Vision
<https://ai.stanford.edu/~syueung/cvweb/tutorial1.html>
- ③ Zdjęcie autorstwa Imansyah Muhamad Putera
<https://unsplash.com/photos/n4KewLKFOZw>
- ④ Zdjęcie autorstwa Isaiah Rustad
<https://unsplash.com/photos/PIhsoerkXxY>
- ⑤ XKCD: Machine Learning <https://xkcd.com/1838/>